T-1
11 Copies

AERONCA
MANUFACTURING CORPORATION

AEROSPACE DIVISION

BALTIMORE 3, MARYLAND

N64-29361	
(ACCESSION NUMBER)	(THRU)
185	1
(PAGES)	(CODE)
CR-56298	06
(NASA CR OR TMX OR AD NUMBER)	(CATEGORY)

FACILITY FORM 602

Downtown [illegible]

185p

Final Report

Contract NASr-25

Aeronca Project 388

31 July 1962

OTS PRICE

XEROX $ 13.00 ph
MICROFILM $ ________

A. Foreword.

In this report it has been attempted to cover the main themes of the effort at producing an optimal control synthesis procedure at Aeronca.

Before describing the presentation, attention should be called to the highlights of this report. In Chapter 3, very recent work actually leading to a "rational" synthesis procedure is discussed. What had previously been a theory largely serving as guidance for a sharper intuition, now for the first time comes into its own as with an algorithm for producing a"closed form"control law. In view of the optimistic comments on the possibilities that this step unfolds to say nothing of the pessimism voiced as to the possibility of ever generating such a "closed form", the chapter deserves special reading. Having actual samples of the "closed form" in hand, brings much more clearly into focus the possibilities - not the least of which is further work - that it affords. In any case, a lot more is known.

Another source of insight will be having a number of trajectories in hand. Work along this line has been directed toward developing a computing procedure that will make these trajectories amenable without undue computing time. This effort is covered in Sections 4.4 and 4.6, where a novel approach taking advantage of the possibility of a steepest descent approach is developed.

Turning now to the body of the report. The first two chapters are given over to developing an understanding of the theory of optimal control. The first chapter discusses the background while the second chapter gives a detailed account of the theory with examples of the generality where it is shown that other criteria than time optimal are also treated by the general theory. In Chapter 3, a theory that actually computes switching surfaces in developed. To our knowledge, these results are presented for the first time here. Chapter 4 gives a presentation of the adjoint system approach as a synthesis and investigation technique. The following chapter covers some preliminary thoughts on the procedure by which the approaches for Chapter 4 might be realized in hardware configurations expressly designed as control computers. In Chapter 6, applications to real plants are discussed. It might be mentioned here that not the least of these applications is to the control of the Saturn, which is a central theme of the work here, and that this work is currently being reported in reports under contract NAS8-5002. Finally there are three appendices being devoted to outlining the algorithm for computing trajectories, giving in order the algebra required, an outline of the steps of a computer program that will effect this algebra and finally a statements of this program in FORTRAN language.

It is to be noted that a wide approach to the synthesis problem has been taken here. Admittedly at this stage a certain disjointedness in the seperate developments exists. To have not broached the problem over this breadth would certainly have led to oversights in the approach, which is to develop a straightforward optimal control synthesis technique. It is expected that as the results flow in the approach will show a more unified viewpoint. Certainly, very material beginnings of this trend are already to be observed.

TABLE OF CONTENTS.

1.0 INTRODUCTION AND BACKGROUND

In order to introduce the subject of time-optimal control, we shall summarize briefly the salient facts about what seems to be the simplest possible example.

Consider an axially symmetric mass, with unit moment of inertia, free to rotate frictionlessly about its axis of symmetry. (This is a fairly good representation of the problem of controlling the roll of a space vehicle). Let θ denote the angle between a radial line fixed in the body, and a radial line fixed in inertial space, and suppose that it is desired to bring the angle θ to zero. (Figure 1.1).

As usual, let $\dot{\theta}$ denote the rate of change of θ, or the angular velocity of the body, i.e.,

(1) $$\dot{\theta} = d\theta/dt.$$

Suppose that it is possible to measure θ and $\dot{\theta}$ at each instant of time t, and suppose that it is desired to apply a controlling torque $\gamma = \gamma(\theta, \dot{\theta})$ in accordance with some prespecified dependence on the instantaneous state $(\theta, \dot{\theta})$ of the body. Suppose finally that themaximum torque available is unity, i.e.,

(2) $$-1 \leqq \gamma(\theta, \dot{\theta}) \leqq +1.$$

Then, according to the well known principles of mechanics, the state $(\theta(t), \dot{\theta}(t))$ of the body will evolve in a manner determined by the differential equation

(3) $$\ddot{\theta} - \gamma(\theta, \dot{\theta}), \; \theta(0) = \theta_o, \; \dot{\theta}(0) = \dot{\theta}_o, \quad |\gamma| \leqq 1.$$

The problem of synthesis of a control system for the body is thus reduced to the choice of the control function $\gamma(\theta, \dot{\theta})$.

One possible control function is

$$(4) \qquad \gamma = -\mathrm{sgn}(\theta + \chi\dot{\theta}), \quad \chi > 0.$$

After a finite number of changes of sign, the system (3-4) reaches a state where the only choice of γ consistent with the physically necessary semi-continuity of $\dot{\theta}$, i.e.,

$$(5) \qquad \dot{\theta}(t + 0) = \dot{\theta}(t),$$

is $\gamma = 0$ for all $t \geqq T_* > 0$. This state $(\theta_*, \dot{\theta}_*)$, first realized by Flugge-Lotz and Klotter [1], is called an end-state; subsequently there are compelling physical reasons, first proved by Andre and Seibert [2], for extending the model (3-5) by the lower-order differential equation

$$(6) \qquad \dot{\theta} + (1/\chi)\theta = 0, \quad \theta(0) = \theta_*, \ \dot{\theta}(0) = \dot{\theta}_* .$$

It can be seen from the example in Figure 1.2 that, in general, several'overshoots' will occur before the motion at last decays exponentially to rest.

Within the past decade it has been realized that it is possible to effect a drastic improvement in the performance of the control system defined by (4). In fact, if the control function

$$(7) \qquad \gamma = -\mathrm{sgn}(\theta + \tfrac{1}{2}\dot{\theta}|\dot{\theta}|)$$

be used, then the body will always come to rest in the desired attitude in a finite length of time T, after at most one change of sign of the controlling torque. Moreover, it has been proved (cf. [3], [4], [5]) that no control function satisfying (2) can bring the body to rest in a time less than T.

This is because the control (7) anticipates the future evolution of the motion and reverses the torque at precisely the time when otherwise unnecessary energy would be added to the body and it would be impossible for a unit torque to halt the motion before at least one overshoot had occurred.

Now virtually all engineering systems of practical interest have more than two degrees of freedom, and obey more complicated electromechanical equations than (3). Hence a theory of time-optimal control for systems with arbitrarily many degrees of freedom and arbitrary governing equations must be developed in order to exploit this concept as a practical system synthesis procedure.

Here we shall attempt to present an introductory exposition to one such general theory, including a statement of the present status of the theory, together with some original results intended to bring the theory to the point of practical effectiveness.

1.1 DEFINITIONS AND STANDING ASSUMPTIONS.

Suppose that the instantaneous state of a certain electromechanical system can be specified by a set of n numbers $x_1, x_2, \ldots, x_n$.

For brevity, we shall use vector notation; thus

$$(1.1) \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} .$$

Henceforth the n-vector x will represent the state of the system in question. The state of the system at time t will be denoted by

(1.2) $$x(t), \quad (-\infty < t < +\infty).$$

We assume that the known laws of physics govern the system's evolution in time, and that these laws lead to an ordinary differential equation of the form

(1.3) $$\dot{x} = F(x), \quad x(0) = x_o, \quad (\cdot = d/dt),$$

where

$$F = \begin{pmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_n(x) \end{pmatrix} = \begin{pmatrix} F_1(x_1, x_2, \ldots, x_n) \\ F_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ F_n(x_1, x_2, \ldots, x_n) \end{pmatrix}$$

is a vector-valued function, or '<u>vector field</u>' defined on the system's state-space.

Under very general conditions, there will correspond to each <u>initial state</u> x_o a solution of (1.3),

(1.4) $$x = X(t;\ x_o), \quad (X(0;\ x_o) = x_o),$$

which is unique. The vector-function X is called the <u>general solution</u> of (1.3); it satisfies the relation

(1.5) $$\partial X(t;\ x_o)/\partial t = F(X(t;\ x_o))$$

identically in the variables t and x_o. If

(1.6) $$X(t;\ x_o) = x_o, \quad (-\infty < t < +\infty),$$

then we call x_o an <u>equilibrium state</u>. Obviously, when (1.6) is inserted into (1.5) one finds that $\partial x_o/\partial t = 0 = F(x_o)$, i.e., that $x = x_o$ is an equilibrium state if and only if

$$(1.7) \qquad F(x_o) = 0.$$

We consider here only those systems which have precisely one equilibrium state. [In a preliminary analysis many systems can be represented by such a mathematical model, at the cost of slight over-idealisation. The effects of such a simplification can be investigated later, after the broad features of the system have been fixed].

By an elementary change of variables, we can assume without loss of generality, and henceforth we do assume, that the equilibrium state has the coordinates

$$(1.8) \qquad x_1 = x_2 = \cdots = x_n = 0,$$

i.e., that it is at the origin $x = 0$ of the state-space.

Thus we confine attention to systems of the form (1.3), subjected to the constraints

$$(1.9) \qquad F(0) = 0,$$

$$(1.10) \qquad F(x) \neq 0 \quad \text{if} \quad x \neq 0.$$

We shall call a dynamical system a control system if it has the property that regardless of its initial state the system's current state always evolves toward the equilibrium state as time increases. In other words, regardless of x_o,

$$(1.11) \qquad X(t; x_o) \to 0 \quad \text{as} \quad t \to +\infty.$$

In mathematical terminology, the system (1.3), (1.9), (1.10) represents a control system if its equilibrium solution $x = 0$ is globally asymptotically stable.

If $F(x)$ is a continuous vector field, then (1.11) represents the best control action that can be obtained, namely, the system tends asymptotically to equilibrium; eventually the system will be arbitrarily near to its equilibrium state, but this does not occur until after the lapse of a corresponding arbitrarily large interval of time.

However, if $F(x)$ is allowed to have discontinuities, then the situation is quite different; in fact, it is then possible for the system's state to reach equilibrium in a finite length of time.

Thus, instead of (1.11), we may consider the possibility that

$$X(T;\ x_o) = 0, \quad T = T(x_o) > 0. \tag{1.12}$$

By the general solution of (3) we now mean a function of the form (1.4) which is, for each fixed x_o, a continuous function of t, and satisfies (5) at all times

$$t_i < t < t_{i+1}, \quad (i = 1,2, \dots), \tag{1.13}$$

where the switching times $\{t_i\}$ form an unbounded monotone increasing sequence

$$0 \leq t_1 < t_2 < \cdots < t_k < t_{k+1} < \cdots ; \tag{1.14}$$

$$t_j \to +\infty \quad \text{as} \quad j \to +\infty. \tag{1.15}$$

Furthermore, denoting $\lim \dot{x}(t_k + \epsilon)$, $\epsilon > 0$, as $\epsilon \to 0$, by $\dot{x}(t_k + 0)$, we require that

$$\dot{x}(t_k + 0) = \dot{x}(t_k), \qquad (k = 1,2, \dots,). \tag{1.16}$$

Thus we allow the graph of the curve $X(t;\ x_o)$ to have "sharp edges", i.e., discontinuities in its tangent line, at the times $t = t_i$, $(i = 1,2,3, \dots)$, and we require "continuity on the right hand side" for this line.

Generally the designer of a control system will have certain elements of the system at his disposal and others which are given and cannot be changed. An extremely common situation is that in which $F(x) = f(x) + Kc(x)$, where the vector field $f(x)$ is given, and where the linear vector function or matrix $K = (K_{ij}; (i = 1, \ldots, n), (j = 1, \ldots, n))$ is also given, but where the control function $c = c(x)$ is not known in advance.

Since no infinite forces or torques can be realized in a macroscopic physical system, the components of the control function will necessarily be subjected to limitations of the form $|c_i(x)| \leqq \alpha_i$, $(i = 1,2,\ldots, n)$, where the constants $\alpha_1, \alpha_2, \ldots, \alpha_n$ are also known in advance.

Without loss of generality, one can assume that each of the bounds $\alpha_i = 1$ [for otherwise replace $c_i(x)$ by $c_i(x)\,\alpha_i$ $(i = 1, \ldots, n)$ and replace K by the matrix KA, where $A = \mathrm{diag}\,(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

1.2 STATEMENT OF THE PROBLEM

STABILITY PROBLEM. Given a control system of the form

$$\dot{x} = f(x) + Kc(x), \; f(0) = 0, \; x(0) = x_o, \tag{1.2.1}$$

where f and K are specified in advance, can one choose a control function $c(x)$, subject only to the constraints

$$0 \leqq |c_i(x)| \leqq 1, \quad (i = 1,2, \ldots, n), \tag{1.2.2}$$

in such a way that the general solution $X(t; x_o)$ is well defined [as in (1.13)-(1.16), except for states x at which $c(x)$ is discontinuous], and satisfies (1.12) for every admissable initial state x_o.

TIME-OPTIMALITY PROBLEM. *Supposing that the answer to the Stability Problem is affirmative, can one choose a particular control function* $c = u(x)$, $|u_i(x)| \le 1$, $(i = 1, \ldots, n)$, *in such a way that, referring to* (1.12), *the transition times* $T(x_o)$ *are always smaller than for any other stable control* [i.e., more precisely, such that

$$0 < T(x_o;\ \{u(x)\}) \le T(x_o;\ \{c(x)\}) \tag{1.2.3}$$

whenever $\{c(x)\}$ is any stable control satisfying (1.2.1), (1.2.2) and 1.12)].

DEFINITION. By the transpose x^* of a column vector

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

we shall denote the *row* vector $x^* = (x_1, x_2, \ldots, x_n)$. By the *transpose* K^* of a square matrix $K = (K_{ij}) = (k_1, k_2, \ldots, k_n)$, where k_i are the *column* vectors of K, we shall denote the matrix

$$K^* = \begin{pmatrix} k_1^* \\ k_2^* \\ \vdots \\ k_n^* \end{pmatrix} \qquad ; \tag{1.2.4}$$

that is, $K^* = (K_{ji})$ where now the rows and columns are transposed. By the *Jacobian matrix* f_x of a vector function f we shall denote the matrix $f_x(x) = (\partial f_i/\partial x_j)$, i.e.,

$$f_x(x) = ([\mathrm{grad}\ f_1(x)]^* , \ldots, [\mathrm{grad}\ f_n(x)]^*)^*, \tag{1.2.5}$$

where f_i are the components of f, and where by the *gradient* $\mathrm{grad}\ \varphi$ of a scalar function $\varphi(x)$ we denote the column vector

(1.2.6)
$$\text{Grad } \varphi = \begin{pmatrix} \partial\varphi/\partial x_1 \\ \partial\varphi/\partial x_2 \\ \vdots \\ \partial\varphi/\partial x_n \end{pmatrix}$$

In case $\varphi = \varphi(x, y)$ then we denote by $\text{grad}_{(x)}\varphi$ and $\text{grad}_{(y)}\varphi$ the results obtained by considering, respectively, y and x as constants.

DEFINITION. The scalar product $x \cdot y$ of two n-vectors is given by

(1.2.7)
$$x \cdot y = x_1 y_1 + x_2 y_2 + \dots + x_n y_n = x^* y = x y^*$$

where the matricial formulation x^*y corresponds to the usual rules of matrix multiplication, namely, if $A = (a_1, a_2, \dots, a_n)^*$ is a matrix whose rows are a_i^*, $(i = 1,2,\dots,n)$, and if b is any n-vector, then

(1.2.8)
$$Ab = (a_1{}^*b, a_2{}^*b, \dots, a_n{}^*b)^* ;$$

and if $B = (b_1, b_2, \dots, b_n)$ is any matrix whose columns are $b_j{}^*$, $(j = 1,2, \dots, n)$, then

(1.2.9)
$$AB = (Ab_1, Ab_2, \dots, Ab_n) = (a_i{}^*b_j).$$

DEFINITION. The length $\|x\|$ of an n-vector x is defined to be

(1.2.10)
$$\|x\| = +\sqrt{x \cdot x} \equiv +\sqrt{x_1{}^2 + x_2{}^2 + \dots + x_n{}^2} \ .$$

DEFINITION. If y is an n-vector, by $\text{sgn}[y]$ we shall denote the vector

(1.2.11)
$$\operatorname{sgn}[y] = \begin{pmatrix} \operatorname{sgn}[y_1] \\ \operatorname{sgn}[y_2] \\ \vdots \\ \operatorname{sgn}[y_n] \end{pmatrix},$$

where by $\operatorname{sgn}[\alpha]$ if α is a real scalar, one means

(1.2.12)
$$\operatorname{sgn}[\alpha] = \begin{cases} +1, & \text{if } \alpha > 0, \\ 0, & \text{if } \alpha = 0, \\ -1, & \text{if } \alpha < 0. \end{cases}$$

DEFINITION. By the exponential e^A of a matrix A one denotes the clearly convergent series

(1.2.13)
$$e^A = \sum_{k=0}^{\infty} A^k/k!$$

where, as usual, $0! = 1$, and where by A^o one denotes the identity matrix

(1.2.14)
$$I_n = \operatorname{diag}(1, 1, \ldots, 1).$$

Note that if A is a constant matrix, the general solution of the linear differential equation

(1.2.15)
$$\dot{x} = Ax, \quad x(0) = x_o$$

is given by $x = X(t; x_o)$ where

(1.2.16)
$$X = e^{tA}x_o;$$

in fact, term-by-term differentiation of (1.2.13) [which can be justified] shows that

(1.2.17) $$(e^{tf})\cdot = Ae^{tA};$$

hence $\partial X/\partial t = Ae^{tA}x_o = Ax$; and the proof is completed by noting that, from (1.2.13) $e^o = I_n$, whence $X(0; x_o) = I_n x_o = x_o$.

1.3 HISTORICAL REMARKS

A special case of the Time-Optimality Problem which has been intensively studied is that in which the given part of the system, $f(x)$, is linear, i.e., $f(x) = Ax$, $A = (A_{ij})$. Then one considers the problem

(1.3.1) $$\dot{x} = Ax + Kc, \quad x(0) = x_o, \quad |c_i| \leq 1, \quad c(0) = 0.$$

This problem (1.3.1) was first studied in the case $n = 2$ by engineers such as MacDonald [6], Hammond and Uttley [7], and Feldbaum [8], and by such mathematicians as Bushaw [4], and LaSalle [5]. A summary of this work appears in Tsien's book [3].

For arbitrary dimensions n the problem (29) was first considered by Rose [9], Lerner [10], Feldbaum [11], and Krassovskii [12]. In 1955, an elegant partial solution to the problem was given by Bellman, Glicksberg and Gross [13]. They first considered the problem of finding c not as a function of x, but as a function of t [that is, they considered "open loop" instead of "closed loop" or feedback control]. Their result (slightly generalized) is that the system

(1.3.2) $$\dot{x} = Ax + K \operatorname{sgn}[K^* e^{-tA^*} y_o], \quad x(0) = x_o$$

is time-optimal whenever

(i) the vectors Ke^i, AKe^i, ..., $A^{n-1}Ke^i$, $(i = 1, \ldots, n)$ are linearly independent, where e^j are the fundamental unit vectors, that is, $I_n = (e^1, e^2, \ldots, e^n)$.

(ii) for every x_o, $\|e^{tA}x_o\| \to 0$ as $t \to +\infty$;

(iii) the vector $y_o = g(x_o)$ is chosen to correspond uniquely with x_o in a certain manner.

The condition (ii) is equivalent to the statement that all eigenvalues of A have negative real parts, which can be ascertained by the Routh-Hurwitz Stability Criteria. Although much work has been devoted to the question (iii) (cf. [12], [14], [15]) and, in principle, effective methods for establishing the correspondence $y_o = g(x_o)$ are known, the practical specification of the function g even for $n = 3$ is a very difficult matter.

Thus, when (i)-(iii) hold, the time-optimal control law for (30) has the form

$$c(t;\, x_o) = \operatorname{sgn}[K^* e^{-tA^*} y_o], \quad y_o = g(x_o). \tag{1.3.3}$$

Now, in general, if $c(t;\, x_o)$ represents the optimal control of (1.2.1) as a function of time, then the optimal control $u(x)$ must satisfy

$$u(X(t;\, x_o)) = c(t;\, x_o), \tag{1.3.4}$$

a statement which Bellman [16], [17] calls the Principle of Optimality. Clearly, then, $u(x_o) = u(X(0;\, x_o)) = c(0;\, x_o)$, and since this holds for arbitrary initial states x_o, one has

$$u(x) \equiv c(0;\, x). \tag{1.3.5}$$

Applying this to the special case (1.3.3), one finds that for (1.3.1) the optimal control is given by

$$u(x) = \mathrm{sgn}[K^*g(x)] \tag{1.3.6}$$

where $g(x)$ *is the same function of* x *that* y_o *is, as a function of* x_o, *in* (1.3.2).

The results (1.3.2) and (1.3.3) are not stated explicitly in matrix notation in [13], and in particular, it is not evident from [13] that $c(t;\, x_o)$ involves e^{-tA^*}. However, in 1955 R. Bass from the work of Bellman, Glicksberg and Gross, noted the results (1.3.3)-(1.3.5) and made the following reformulation. *Consider the system*

$$\dot{x} = Ax + K\,\mathrm{sgn}[K^*y], \quad x(0) = x_o \tag{1.3.7}$$

and its "adjoint system"

$$\dot{y} = -A^*y,\ y(0) = y_o, \tag{1.3.8}$$

as a simultaneous 2n-dimensional system. If the function $g(x)$ *has the property that for each* $x_o \neq 0$ *there is a* $T = T(x_o) > 0$ *such that*

$$\underline{\text{if}}\ \ y_o = g(x_o)\ \ \underline{\text{then}}\ \ X(T(x_o);\, x_o) = 0 \tag{1.3.9}$$

then the system

$$\dot{x} = Ax + K\,\mathrm{sgn}[K^*g(x)], \quad g(0) = 0 \tag{1.3.10}$$

is the time-optimal.

These published results (1.3.7),(1.3.8),(1.3.9),(1.3.10) in 1956 [18] *explicitly* suggest that the *adjoint system* (1.3.8) can be used for numerical tabluation of the optimal control law $g(x)$ as follows.

Simulate the system (1.3.7)-(1.3.8) by means of analog or digital computer. For each fixed x_o, vary y_o until, by trial-and-error, a value g_o is found such that with $x(0) = x_o$, $y(0) = g_o$, there is a $T > 0$ such that $X(T, x_o, g_o) = 0$. Then repeat for a different x_o. In this way a table of pairs (g_o, x_o) can be constructed. But such a table defines a function $g(x)$ such that $g(x_o) = g_o$ for every x_o; and then (1.3.6) gives the time-optimal control law.

To be sure, the preceding prescription represents a formidable task - clearly unfeasible for very large value of n. Nevertheless, drastic reductions in the amount of computing can be made, as will be indicated later. Since 1956, well-known classical results about the Hamiltonian formulation of the Problem of Bolza, the Weierstrass E-Function, and Hilbert's Integral, and in particular the necessary maximality of the Hamiltonian as a function of c have come to light. These lead at once to (1.3.7)-(1.3.8) [when trivial modifications to allow for constraints of the form $|c_i| \leq 1$ rather than $|c_i| < 1$ are made in the classical statements; cf., e.g., Caratheodory's book [19] and Breakwell's paper [20]].

Independently, the same problem was considered in the USSR by Pontrjagin and his collaborators Gamkrelidze, Boltianskii, and Mishchenko. In 1956 Pontrjagin announced [22] a conjecture regarding the time-optimality problem, and in 1957 and 1958 the conjecture was proved true ([23], [24], [25]) under conditions of great generality. Their result, called the Maximum Principle, contains the preceding results (1.3.7)-1.3.10) as a special case, and generalizes them to a wide class of nonlinear systems.

1.4 FORMULATION OF THE MAXIMUM PRINCIPLE

DEFINITION. By E^n, or n-dimensional Euclidean space, we denote the set of all n-vectors, $x, y, \ldots,$ considered as the radius vectors of points, with the distance between two points defined by means of the metric

$$\rho(x,y) = \|x-y\|, \tag{1.4.1}$$

where by $\alpha x + \beta y$, for arbitrary numbers α and β, one denotes the vector with components $\alpha x_i + \beta y_i$. The set of all points having radius vectors x such that $\rho(x, x_o) = \|x-x_o\| < R$ is called the *spherical neighborhood of* x_o *of radius* R. A subset G of E^n is *open* if each point g in G has some spherical neighborhood [so small that it can be] entirely contained in G. A subset F is *closed* if its *complement* $G = F^c$ [i.e., the set of all points of E^n not in F] is open. A set D is *bounded* if it is contained in some [sufficiently large] spherical neighborhood of the origin.

THE MAXIMUM PRINCIPLE. *Consider the nonlinear dynamical system*

$$\dot{x} = f(x, c), \quad x(0) = x_o, \quad f(0, 0) = 0 \tag{1.4.2}$$

$$x \ \underline{\text{in}}\ G, \quad c \ \underline{\text{in}}\ C \quad ((0,0)\ \underline{\text{in}}\ G\ C). \tag{1.4.3}$$

where G *is an open subset of* E^n, *and* C *is a closed and bounded subset of* E^n; *we suppose that both* f_x *and* f_c *are continuous functions of* (x, c). *Suppose that it is desired to choose the control function* $c = c(x)$ *in such a way that it satisfies the constraint (1.4.3) and that the general solution of (1.3.2),* $X = X(t; x_o; \{c(x)\})$, *is defined for all values of* x_o *at which* $c(x_o)$ *is continuous, and, for each fixed such* x_o, *is a continuous piece-wise differentiable function of* t *which satisfies* $\partial X/\partial t = f(X, c(X))$ *for all values of* t *at which* $c(X(t; x_o))$ *is continuous, and which, for every admissible* x_o *satisfies*

$$X(T; x_o; c(x)\ ) = 0 \tag{1.4.4}$$

for some number $T = T(x_o; \{u(x)\}) > 0$. *If now there is a time-optimal control function* $c = u(x)$ *such that*

(1.4.5) $$0 < T(x_o; \{u(x)\}) \leq T(x_o; c(x))$$

for all admissible control functions $\{c(x)\}$, *then* $\{u(x)\}$ *must satisfy the following conditions. Define*

(1.4.6) $$c_*(t; x_o) = u(X(t; x_o; \{u(x)\})$$

define the scalar function

(1.4.7) $$\varphi = \varphi(x, y; c) \equiv y \cdot f(x, c);$$

and consider the Hamiltonian system

(1.4.8) $$\dot{x} = \text{grad}_{(y)}\varphi \Big|_{c = c_*} = f(x, c_*(t; x_o)), \quad x(0) = x_o$$

(1.4.9) $$\dot{y} = -\text{grad}_{(x)}\varphi \Big|_{c = c_*} = -f_x{}^*(x, c_*(t, x_o))y, \quad y(0) = y_o \ .$$

Then for each admissible x_o *there is a* $y_o = g(x_o)$ *such that the general solution* $X(t; x_o)$ *of (1.4.8) satisfies*

(1.4.10) $$X(T, x_o) = 0$$

for some $T > 0$ *and such that*

(1.4.11) $$\varphi_o = \varphi(x_o, y_o; c(0; x_o)) \geq 0$$

while for $0 \leq t \leq T = T(x_o)$

$$\varphi(x(t), y(t); c_*(t, x_o)) = \Phi(x(t), y(t)), \tag{1.4.12}$$

where, by definition, for all x *in* G *and* y *in* E^n

$$\Phi(x,y) = \max_{c \text{ in } C} \varphi(x, y; c). \tag{1.4.13}$$

Furthermore, for $0 \leqq t \leqq T(x_o)$,

$$\varphi(x(t), y(t); c_*(t, x_o)) = \varphi_o \geqq 0. \tag{1.4.14}$$

Finally, as a consequence of (1.4.6) and (1.4.12), one has that, for $y_o = g(x_o)$,

$$\varphi(x(t), y(t); u(x(t)) = \Phi(x(t), y(t)), \tag{1.4.15a}$$

and, in particular,

$$\varphi(x_o, y_o; u(x_o)) = \Phi(x_o, y_o). \tag{1.4.15b}$$

COROLLARY. *If for each admissible* x_o *one can solve the boundary-value problem (1.4.8), (1.4.9), (1.4.10), (1.4.11), then the correspondence between the initial state* x_o *and its conjugate initial state* $y_o = g(x_o)$ *defines a function* $g(x)$ *which must satisfy the Maximum Principle*

$$\varphi(x, g(x); u(x)) = \Phi(x, g(x)); \tag{1.4.16a}$$

that is,

$$g(x) \cdot f(x, u(x)) = \max_{c \text{ in } C} g(x) \cdot f(x, c) \tag{1.4.16b}$$

Frequently, the principle (1.4.16) enables one to determine the optimal control function $u(x)$ in a unique manner.

For example, consider the time-optimality problem for (1.2.1), (1.2.2). Here $f(x, c) = f(x) + Kc$ and so by (1.4.15a) and (1.4.16a)

(1.4.17)
$$y \cdot f(x) + y \cdot Ku = \max_{|c_i| \leq 1} y \cdot f(x) + y \cdot Kc,$$

that is,

(1.4.18)
$$K^*y \cdot u = \max_{|c_i| \leq 1} K^*y \cdot c = \sum_{i=1}^{n} |[K^*y]_i|$$

which is obtained by the choice $c_i = \mathrm{sgn}[K^*y]_i$. Hence

(1.4.19)
$$\sum_{i=1}^{n} u_i [K^*y]_i = \sum_{i=1}^{n} [K^*y]_i|, \ -1 \leq u_i \leq 1,$$

which implies that, for $y_o = g(x_o)$,

(1.4.20)
$$u = \mathrm{sgn}[K^*y].$$

Similarly, use of (1.4.16b) implies that

(1.4.21)
$$u(x) = \mathrm{sgn}[K^*g(x)], \quad (g(0) = 0)).$$

Thus we may state the following result, discussed below.

REFORMULATION OF THE TIME-OPTIMALITY PROBLEM. *Consider the boundary-value problem defined by the simultaneous systems*

(1.4.22) $$\dot{x} = f(x) + K \operatorname{sgn}[K^*y], \quad x(0) = x_o, \quad x(T) = 0,$$

(1.4.23) $$\dot{y} = -f_x(x)y, \quad y(0) = y_o, \quad \|y_o\| = 1.$$

Suppose that for each $x_o \neq 0$ *there is a* $T = T(x_o) > 0$ *and a corresponding unique* $y_o = g(x_o)$ *such that (1.4.22) and (1.4.23) are satisfied. Then the optimal control law for the system (1.2.1) is given by (1.4.21).*

NOTE. If, for some particular index j, $[K^*(x_1)g(x_1)]_j = 0$, it may be seen by inspection of (1.4.17)-(1.4.21) that any value, subject to the condition $-1 \leq u_j(x_1) \leq 1$, may be selected for $u_j(x_1)$. In this case, the time-optimal control function is not unique. Additional optimality criteria may be imposed to find the most preferable time-optimality control.

1.5 RESOLUTION OF THE PROBLEM

In summary, we may state the following.

RESOLUTION OF THE TIME-OPTIMALITY PROBLEM. *If the function* $g(x)$ *is defined by the correspondence* $y_o = g(x_o)$ $g(0) = 0$ *between the initial states* (x_o, y_o) *which uniquely satisfy the equations (1.4.22 and (1.4.23) then*

(1.5.1) $$\dot{x} = f(x) + K \operatorname{sgn}[K^*g(x)]$$ *is a time-optimal control system.*

REMARK. If, in (1.2.1) one allows $K = K(x)$ instead of K = constant, then all of the statements concerning (1.4.22) (1.4.23) and (1.5.1) remain valid provided that one replaces the adjoint system $\dot{y} = -f_x{}^*(x)y$ by

(1.4.23 bis)
$$\begin{aligned}\dot{y} &= -[y\cdot f(x) + y\cdot K(x)c]_x \Big|_{c = \operatorname{sgn}[K^*(x)y]} \\ &= -[y\cdot f(x) + K^*(x)y\cdot c]_x \Big|_{c = \operatorname{sgn}[K^*(x)y]} \\ &= -f_x^*(x)y - ([K^*(x)y]_x)^*\operatorname{sgn}[K^*(x)y]\end{aligned}$$

Incidentally, in practice it is far more convenient to define the adjoint system as in the first step of (1.4.23 bis), and then to carry through the indicated calculations in each special case, than to calculate the Jacobian matrix $f_x(x)$ and thento transpose it. In fact, for large n the matrix $f_x(x)$ may have zeros in almost every position, and keeping track of them can be extremely tedious; whereas taking the gradient of $\sum_{i=1}^{n} y_i f_i(x)$ is very efficient.

Similarly, it is far more convenient to calculate the gradient of $\sum_{i=1}^{n} y_i [K(x)c]_i$ than to use the general formula $([K^*(x)y]_x)^*$. Some examples of this will be given elsewhere [26], [27].

It should also be noted that the Maximum Principle, as proved in [25], is a necessary but not a sufficient criterion. That is, if one calls any control function obtained by the principle an <u>extremal control</u>, then the theorem of [25] asserts only that <u>every optimal control must be an extremal control</u>.

If, however, there is precisely one extremal control, then there are just two possibilities:

(a) there is no optimal control; or

(b) the extremal control is optimal.

In this way one can,in practice, often by-pass the difficult question as to whether or not an optimal control exists. In fact, if the adjoint-system approach establishes that there is precisely one extremal control function, then this function will certainly provide a stable control; then this function will certainly provide a stable control; and the practical suitability of such a control can be investigated by a computer simulation. To be sure, the theoretical possibility exists that the extremal control does not provide time-optimal control, but rather only the analog of a flex-point of a curve or a saddle-point of a surface, i.e., a "time-stationary" but neither time-minimal nor time-maximal control.

Further study of this question is clearly indicated. It appears likely that under suitable hypotheses the Maximum Principle provides conditions which are not only necessary but also sufficient for optimality. In fact, in 1959 Breakwell [20] published a sufficient condition for the existence of optimal control which is extraordinarily similar to the Maximum Principle; however, in his formulation a problem is degenerate if, for example, the matrix $K = (k_1, k_2, \ldots, k_n)$ in (1.2.1) has more than one non-zero element in any one of its constituent column vectors k_i, $(i = 1, 2, \ldots, n)$. Since this restriction rules out some of the most important practical problems known, an investigation of the possibility of extending Breakwell's sufficient conditions would be most desirable.

REFERENCES

1. I. Flugge-Lotz, *Discontinuous Automatic Control*, Princeton University Press, (1953).

2. J. Andre and P. Seibert, Uber stuckweise lineare Differentialgleichungen, die bei Regelungsproblemen auftreten, I., II., *Archiv der Mathematik*, vol. 7(1956), pp. 148-156 and 157-164.

3. H. S. Tsien, *Engineering Cybernetics*, New York, McGraw-Hill(1954).

4. D. W. Bushaw, Optimal Discontinuous Forcing Terms, *Contributions to the Theory of Nonlinear Oscillations*, vol. IV(1953), Princeton University Press, pp. 29-52.

5. J. P. LaSalle, Basic Principle of the Bang-Bang Servomechanism, Bulletin, AMS, vol. 60(1954), p. 154.

6. D.C. McDonald, Intentional Nonlinearization of Servomechanisms, *Proceedings*, *Symposium on Nonlinear Circuit Analysis*, MRI Symposia Series, vol. 2(1953), Polytechnic Institute of Brooklyn, pp. 402-411.

7. A. M. Uttley and Ph. H. Hammond, The Stabilization of On-Off Controlled Servomechanisms, *Automatic and Manual Control*, London, Butterworths, 1952, pp. 285-307.

8. A. A. Feldbaum, On the Design of Optimal Systems by Means of Phase Space, *Automatika i Telemechanika*, vol. 16(1955), pp. 129-149.

9. N. J. Rose, Theoretical Aspects of Limit Control, *ETT Report Number 459*, Stevens Institute of Technology, Hoboken(1953).

10. A. Y. Lerner, On the limit of high-speed systems of automatic regulation, *Automatika i Telemekhaniki*, Moscow, 15, no. 6(1954), p. 461.

11. A. A. Feldbaum, *Automatika i Telemekhanika*, vol. XIV, no. 6(1953), pp. 212-228.

12. N. N. Krassovskii, On the Theory of Optimal Regulation, *Journal for Applied Mathematics and Mechanics*, vol. XXIII(1959).

13. R. Bellman, I. Glicksberg, O. Gross, On the Bang-Bang Control Problem, *Quarterly of Applied Mathematics*, vol. 13,(1955), pp. 321-324.

14. R. V. Gamkrelidze, Towards a theory of optimal processes and linear systems, *Doklady Akademii Nauk SSSR*, vol. 116, No. 1, (1957).

15. J. P. LaSalle, Time Optimal Control Systems, *RIAS*(Baltimore) Monograph 59-4.

16. R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, N. J. (1957).

17. R. E. Bellman, I. Glicksberg, O. Gross, Some Aspects of the Mathematical Theory of Control Processes, *USAF Project, RAND Corporation*, January, 1958.

18. R. W. Bass, Equivalent linearization, nonlinear circuit synthesis, and the stabilization and optimization of control systems, Proceedings, Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, vol. VI(1956), pp. 163-198.

19. C. Caratheodory, Variationsrechnung und Partielle Differentialgleichungen Erster Ordnung, vol. 1(1956), B. G. Teubner.

20. J. V. Breakwell, The Optimization of Trajectories, Journal, SIAM, vol. 7(1959), pp. 215-247.

21. L. I. Rozonoer, L. S. Pontrjagin's Maximum Principle in the theory of Optimum Systems, Part I, Automatika i Telemekhanika, vol. 20, no. 10(October 1959), pp. 1320-1334.

22. V. G. Boltyanskii, R. V. Gamkrelidze, L. S. Pontrjagin, On the theory of optimum processes, Doklady Akademii Nauk, SSSR, vol. 110, (1956).

23. V. G. Boltyanskii, The maximum principle in the theory of optimal processes, Doklady Akademii Nauk, SSSR, vol. 119, no. 6(1958), pp. 1070-1073.

24. R. V. Gamkrelidze, On the general theory of optimal processes, Doklady Akademii Nauk, SSSR, vol. 123, no. 2 (November 1958).

25. L. S. Pontrjagin, V. G. Boltyanskii, R. V. Gamkrelidze, Optimal Control Processes, Uspekhi, Matematicheskikh Nauk, vol. 14, no. 1(1959), pp. 3-20.

26. P. A. Castruccio, R. W. Bass, D. L. Slotnick, The optimal space vehicle attitude-control system for arbitrary angular variations including nonlinear coupling, Aeronca Astromechanics Institute Technical Report No. 60-23.

27. P. A. Castruccio, R. W. Bass, D. L. Slotnick, Feedback computer for optimal thrust-vector attitude control and propulsion switching in mid-course and landing maneuvers, Aeronca Astromechanics Institute Technical Report No. 60-24.

2.0 THEORETICAL FOUNDATION OF OPTIMAL CONTROL SYSTEM SYNTHESIS.

2.1 The Variational Approach as a Formulation of the Synthesis Problem.

(1) System Dynamics. The evolution of the dynamical system, represented by the time dependent state vector or curve $x(t)\epsilon E^n$ is assumed to be determined by the differential system

$$(1) \qquad \dot{x} = f(x, u(x)),$$

with

$$x(0) = x_0$$

where x_0 is the initial state and $u(x)\epsilon E^n$ the control function. (If the vector u depends only on the state x, as defined here, u is called "instantaneous state feedback "control function") .

(2) a. The class $\mathcal{U}$ of admissible control functions is defined by means of an open*, bounded, convex subset of U in E^n. It is constituted by continuous and continuously differentiable functions $u(x)$ in some subset R of E^n. Explicitly

$$(2) \qquad \mathcal{U} \triangleq \left\{u(x) \,\middle|\, x\epsilon R,\quad u\epsilon U,\quad u(x)\epsilon C^0,\ D',\quad u(0) = 0,\right\}$$

where: $R \subset E^n$, R open, $\{x = 0\}\epsilon R$, and

$U \subset E^n$, U open and bounded, convex

$C^0 \triangleq$ class of continuous functions

*This means that there is no saturated (discontinuous) control c_s, as we are restricted here to continuous control ($c_s\epsilon\partial U$, or boundary of U, only).

$C^1 \triangleq$ class of continuously differentiable functions everywhere in $R \subset E^n$

$D' \triangleq$ class of functions continuously differentiable in an open dense subset R^o of R, i.e. a set of R^o such that $R^o \subset R$ and $\overline{R^o} \cap R = R$.

The condition for the range of control function to belong to a certain set

$$u \in U, \text{ for all } x \in R$$

is called the "control constraint".

We shall distinctly say either $u \in \mathcal{U}$ or $u \in U$.

b. Controllability and stability domain: the dynamical system (1) is assumed to be controllable in R, that is, $\mathcal{U}$ is non-empty (with respect to R) so that for any $x_o \in R$, there exists some time T and at least one of the $u \in \mathcal{U}$ such that

$$x(T) = 0.$$

The region R is called the stability domain of the system with respect to the class of control functions $\mathcal{U}$, and T is the transition time between the states $x_o(0)$ and $x = 0$.

(3) Liapunov function and global asymptotic stability.

(a) Let $f(x,u)$ be a vector function continuously differentiable everwhere in R.

$$(3) \qquad f: R \times U \longrightarrow E^n, \quad f(x,u) \in C'$$

with the property

$$f(0,0) = 0.$$

We shall first assume that a Liapunov function relative to $f(x,u)$ and to the class $\mathcal{U}$ exists, according to the following hypothesis on $\mathcal{U}$ and definition.

Hypothesis on $\mathcal{U}$. For each $u \in \mathcal{U}$ there exists a real function $\rho(x)$ on $R \subset E^n$ with the properties:

$$(4)\quad \begin{array}{ll} \text{(i)} & \rho(0) = 0, \quad \rho(x) > 0 \text{ if } x \neq 0 \\ \text{(ii)} & \rho(x) \longrightarrow +\infty \text{ as } x \longrightarrow \partial R \text{ if } R \text{ is bounded, or} \\ & \rho(x) \longrightarrow +\infty \text{ as } \|x\| \longrightarrow \infty \text{ if } R \text{ is unbounded} \\ \text{(iii)} & \text{the Lie derivative of } \rho(x) \text{ is negative definite, i.e.} \\ & f(x,u(x))\cdot \text{grad } \rho(x) < 0, \text{ for } x \neq 0 \\ & \qquad\qquad\qquad = 0, \text{ for } x = 0. \end{array}$$

<u>Definition 2.1</u>. A function $\rho(x)$ defined on $R \subset E^n$ with the properties (i), (ii), (iii) is called a <u>Liapunov</u> <u>function</u> relative to f and to the class of controls $\mathcal{U}$.

<u>Lemma 1</u>. Consider the vector function $f(x,u) \in C^1$ such that $f(0,0) = 0$, as defined above. For each $u \in \mathcal{U}$, there exists a vector function $F(x,t) \in C^1$.

$$F\colon\ R\mathrm{x}I_+ \longrightarrow E^n$$

for all $x \in R$ and $t \in I_+$ $(I_+ \triangleq [0, +\infty])$, with the properties:

$$\text{(i)}\quad \frac{\partial F(x,t)}{\partial t} = f(F(x,t),\ u(F(x,t))),$$

$$\text{(ii)}\quad F(x,0) = x \text{ for every } x \in R,$$

$$\text{(iii)}\quad F(F(x,t_1),t_2) = F(x,t_1 = t_2), \text{ for every } x \in R \text{ and } t_i \in I_+,$$

$$\text{(iv)}\quad F(x,t) \longrightarrow 0 \text{ as } t \longrightarrow +\infty, \text{ for every } x \in R.$$

<u>Proof</u>. It follows immediately from properties (4) of $\rho(x)$.

It is now immediate, as a corollary:

<u>Lemma 2</u>. For any $x_o = 0$, $x_o \in R$, and $u(x) \in \mathcal{U}$, there exists in R a vector functions

$$x(t) = F(x_o,t) \quad (F \text{ defined as in Lemma 1})$$

with the properties

$$\text{(i)}\quad \dot{x} = f(x,u(x)), \quad x(0) = x_o \qquad (\dot{} = d/dt)$$

$$\text{(ii)}\quad x(t) \neq 0, \quad t \in I_+$$

$$\text{(iii)}\quad x(t) \longrightarrow 0 \text{ as } t \longrightarrow +\infty.$$

In other words, $F(x,t)$ is the <u>general</u> <u>solution</u> of $\dot{x} = f(x,u(x))$ with the arbitrary initial condition $x(0) = x_o \in R$, as is <u>globally asymptotically stable</u> in R with respect to $\mathcal{U}$.

<u>Proof</u>: It follows from Lemma 1.

(b) Construction of a Liapunov function. We shall now heuristically "construct" a Liapunov function after the following:

Definition 2.2. For every $u \in \mathcal{U}$, define a non-negative real function $\alpha(x,u)$

$$\alpha : R \times U \rightarrow I_+$$

with the properties

(i) $\alpha \in C^1$

(ii) $\alpha(x,u) > 0$ for $(x,u) \neq (0,0), \in R \times U$

$\alpha(0,0) = 0$

(iii) $\mathrm{grad}_{(x)}\,(0,u) \equiv 0$ (identically with respect to U).

One can always assume that such a function α exists.

Definition 2.3 (Performance Index). (a) For each $x \in R$ and $u \in \mathcal{U}$ define a real function $\varphi : R \rightarrow I_+$ by the following integral, which is in fact a functional:

$$\varphi(x_o) \triangleq \int_0^{\infty} \alpha\left[x(t),\ u(x(t))\right] dt$$

where $\alpha(x,u)$ is a definite positive real function as defined in Def. 2.2. This function is called "performance index" or "performance criterion" in control theory. (b) Accordingly, and by the definition of $x(t) = F(x_o,t)$, the φ above is equivalent to

$$\varphi(x) \triangleq \int_0^{\infty} \alpha\left[F(x,t),\ u(F(x,t))\right] dt$$

(where, since F was defined for every $x_o \in R$, x_o and x have been interchanged).

Hypotheses on φ.

1. The integral defining $\varphi(x)$, for every $u \in \mathcal{U}$, converges and thus φ exists.

2. $\varphi \in D^1$ and therefore grad φ exists in $R^o \subset R$ (i.e. piecewise in R)[1]. We shall

[1] More generally one could state that φ is smooth almost everywhere (a.e.) in R and thus grad φ exists a.e. in R. Under the more restrictive hypothesis 2 used here, φ is said to be "Regular in R".

prove now that $\varphi(x)$ so defined is a Liapunov function.

Lemma 3. For each $u(x) \in \mathcal{U}$. $\varphi(x)$ of Def. 2.3 is a Liapunov function (Def.2.1).

Proof. First, it is clear that, by definition, φ is positive definite with $\varphi(0) = 0$, which proves conditions (4) (i) and (ii). Second, let us prove (4) (iii), i.e., the Lie derivative of $\varphi(x)$ is negative definite. For this purpose, it is sufficient to prove an even more particular property which characterizes $\varphi(x)$, namely

$$(5) \qquad f(x,u(x)) \cdot \mathrm{grad}\, \varphi(x) \equiv -\alpha(x,u(x))$$

for every $x \in R$, with $\alpha(x,u) \geq 0$ as defined.

Proof. If $F(x,t)$ is the general solution of $\dot{x} = f(x,u(x))$, then (calling $\sigma \in I_+$ the independent variable),

$$\varphi(F(x,t) = \int_0^{+\infty} \alpha\Big[F(F(x,t),\sigma), u(F(F(x,t),\sigma))\Big] d\sigma,$$

which holds for any $x \in T$.

By (iii) of Lemma 1, the same can be written

$$\varphi(F(x,\sigma)) = \int_0^{+\infty} \sigma\Big[F(c,\sigma+t), u(F(x,\sigma+t))\Big] d\sigma.$$

Setting $\Theta = \sigma + t$, $d\sigma = d\Theta$, with $t \leq \Theta \leq +\infty$, and again (writing σ instead of Θ):

$$\varphi(F(x,\sigma) = \int_t^{+\infty} \alpha\Big[F(x,\sigma), u(F(x,\sigma))\Big] d\sigma.$$

The derivative of φ with respect to σ is then (writing again t instead of σ),

$$\frac{d\varphi(F(x,t))}{dt} = \frac{\partial F(x,t)}{\partial t} \cdot \operatorname{grad} \varphi(F(x,t)) =$$

$$= f(F(x,t), u(F(x,t)))\cdot \operatorname{grad} \varphi(F(x,t)) =$$

$$= -\alpha(F(x,t), u(F(x,t)))$$

for $t\epsilon I_+$ and $x\epsilon R$, as follows from (i) of Lemma 1 and from "Hypothesis on $\mathcal{U}$", (4) (iii). Setting $t = 0$, by (ii) of Lemma 1 stating that $F(x,0) \equiv 0$, the identity (5) follows, i.e. the Lie derivative of φ is negative definite,

(6) $$\begin{aligned} f(x,u(x))\cdot \operatorname{grad} \varphi(x) &< 0 \text{ for } x = 0 \\ &= 0 \text{ for } x = 0, \end{aligned}$$

hence, by definition, $\varphi(x)$ is a Liapunov function relative to the class $\mathcal{U}$. Q.E.D.

The control problem involves the following: We are given a smooth control system of the state vector form

(7) $$\dot{x} = f(x,u) \qquad x(0) = x_0\epsilon R \qquad (\dot{\ }=d/dt)$$

where

(8) $$f(0,0) = 0$$

and we desire to discover a smooth function $u = u(x)$ such that, for some open subset U of E^n,

(9) $$u\epsilon U \quad \text{for all} \quad x\epsilon R,$$

and such that

(10) $$u(0) = 0,$$

and

(11) $$x(t) \to 0 \quad \text{as} \quad 0 \leq t \to +\infty .$$

We have called such a $u(x)$ an admissible control law. It can be shown that a control $u \epsilon U$ law is admissible if and only if there exists on R a positive definite function $\varphi(x)$, and on $R x U$ a positive definite function $\alpha(x,u)$, i.e.

(12) $$\varphi(x) > 0, \quad x \neq 0; \quad \varphi(0) = 0;$$

(13) $$\alpha(x,u) > 0, \quad (x,u) \neq 0; \quad \alpha(0,0) = 0,$$

such that φ is a Liapunov function for (1) and $-\alpha$ is its Lie derivative, i.e.

(14) $$f(x,u(x)) \cdot \operatorname{grad} \varphi(x) = -\alpha(x,u(x))$$

identically on the domain of stability R. In fact, we assume that such a positive definite function φ exists. (Cf. Hypothesis on φ).

As an immediate consequence of (14), we have that

(15) $$\varphi(x_o) = \int_0^{+\infty} \alpha(x(t), u(x(t)))dt.$$

Conversely, if we are given an admissible control $u(x)$, and a performance index $\alpha(x,u)$ which defines a performance criterion φ as in (15), then the convergence of the integral in (15) for all x_o in R implies that (14) holds in R.

The variational approach consists in determing a function $u(x)$ which minimizes the functional $\varphi(x)$ over R.

Definition 2.4. Optimal control law $c(x)$ is a single valued vector function of the class $\mathcal{U}$ of "admissible" controls such that it minimizes (absolute minimum) the functional $\varphi(x) \geq 0$ (Def. 2.3)

$$\varphi(x_o) = \int_0^\infty \alpha(x,c(x))dt = \text{Min.}, \tag{16}$$

(φ defined for any $x_o \epsilon R$) with respect of all the control functions of the class. This property is also expressed by saying that

$$\underline{\varphi} \triangleq (x;\{c\}) \leq \varphi(x;\{u\}) \tag{17}$$

for all $u \epsilon \mathcal{U}$ and $x \epsilon R - 0$. Let us denote by $\{\varphi\}$ the family of al $\varphi(x;\mathcal{U})$.

First Fundamental Hypothesis. It is assumed that $c(x)$ exists (and thus $\underline{\varphi}$, the absolute minimum relative to $\mathcal{U}$, exists).*

Auxiliary hypothesis on $\{\varphi\}$. The minimum $\underline{\varphi}$ is attained (relative to $\mathcal{U}$).

Definition 2.5 Conjugate state or co-state of a system is a single valued function $y\colon R \rightarrow E^n$,

$$y = y(x) \triangleq -\text{grad } \varphi(x), \tag{18}$$

defined piecewise in R, i.e., in $R^o \subset R \subset E^n$ (assuming that $\varphi(x)$ is smooth in $R^o \subset R$ and hence grad $\varphi(x)$ exists in $R^o \subset R$). One also says that y is the

*We do not assume that $c(x)$ is unique. It is clear, however, that $\underline{\varphi}$ (absolute minimum) is unique.

co-state of the state $x(t)$.

Definition 2.6. *The Hamiltonian* of a dynamical system is the single valued real function

$$(19) \qquad H(x,y,u) \triangleq y \cdot f(x,u) - \alpha(x,u)$$

defined on $R x E^n x U$ (with $\alpha(x,u)$ as per Def. 2.2).

Second Fundamental Hypothesis. The Hamiltonian H has a maximum with respect to all control vectors $u \in U$

$$(20) \qquad H(x,y) = \max_{u \in U} H(x,y,u)$$

for each fixed state x (and an arbitrarily fixed co-state y).

Definition 2.7. Define a vector function $c(x,y)$ for every y E^n, such that for all y's, $\bar{c}(x,y) = c \in U$ and such that

$$(21) \qquad H(x,y,\bar{c}(x,y)) = \bar{H}(x,y)$$

Remark. $\bar{c}(x,y)$ is not necessarily single-valued.

2.2 Pontrjagin's Maximum Principle

The statement that $u(x)$ is an admissible control law is equivalent to the statement . which is equivalent to the statement

$$(1) \qquad H(x,y(x), u(x)) \equiv 0.$$

Hence we may use as a mnemonic the remark that "a control system is stable if $H = 0$."

Now consider the choice of an optimal control function $c(x)$, i.e., one which minimizes $\varphi(x_o)$ for each x_o in R. Define

$$(2) \qquad \bar{H} = \bar{H}(x,y) = \max_{u \in U} H(x,y,u).$$

Also, define $\bar{c} = \bar{c}(x,y)$ as any function (in general, it may be multiple-valued as stated before) which satisfies

$$(3) \qquad \bar{H}(x,y) = H(x,y,\bar{c}(x,y)).$$

In practice, it is easy to compute $\bar{c}(x,y)$; one can usually solve

$$(4) \qquad \operatorname{grad}_{(u)} H(x,y,\bar{c}) = 0$$

quite explicitly for $\bar{c} = \bar{c}(x,y)$, and verify that this $\bar{c}$ actually maximizes H by the usual methods of calculus. Let us consider the following Lemma without proof.

<u>Lemma 4.</u> If $c(x)$ that minimizes $\varphi(x)$ exists (Def. 2.4) and the minimum $\bar{H}(x,y(x))$ also exists (for all y), then

(5) $$H(x,y(x),\ c(x) = H(x,y(x)) \equiv 0$$

for every $x \in R = 0$.

<u>Lemma 5.</u> If there exits an optimal control function $c(x)$ (Def. 4.1) and an absolute minimum $\bar{H}(x,y(x))$ of the Hamiltonian, then for every $x \in R$,

$$\text{grad}_{(u)}\alpha(x,u(x))\ _{u=c(x)} + f^*_u(x,c(x))\cdot\text{grad}\ \varphi(x;\{c\}) \equiv 0,$$

<u>Fundamental Theorem of the Maximum.</u>

Let R E^n be an open set and $x = 0 \in R$, and U an open, bounded, and convex set of E^n which also contains the origin $u = 0$ (Cf. 2.3).

Let, as established in Part 3(a), $f\colon\ R \times U \to E^n$ be of class C and $f(0,0) = 0$. Let a real single valued function $H\colon\ R \times E^n xu \to E'$ be defined by $H(x,y,u) = y\cdot f(x,u) - \alpha(x,u)$, with $y = y(x) = -\text{grad}\ \varphi$. With these definitions: <u>Hypotheses</u>: (1) Let consist of the non-empty class of all vector functions $: R \to U$ such that (i) $u(0) = 0$, (ii) $u(x) \in C^o$ on R, (iii) the Jacobian $u_x(x)$ exists and is continuous on an open, dense subset R^o of R. (i.e., there exist controls with property (iii)). (2) There exists a real function $\varphi(x)$ which can be defined as

$$\varphi(x_o) = \int_o^{+\infty} \alpha(x(t),\ u(x(t)dt$$

which is such that $\varphi(0) = 0$, $\varphi(x) > 0$ if $x \neq 0$.

That is, combining (1) and (2), let be the class of "control" functions $u(x)$ and $\varphi(x)$ a Liapunov function, such that ensures the global asymptotic stability of the differential system (1);

$$\dot{x} = f(x, u(x)) \qquad x(0) = x_o$$

for every x_o R $(x_o \neq 0)$.

(3) Suppose that $\varphi(x)$ has an absolute minimum and there exists a function $c(x) \in$ such that

$$\varphi(x; \{c\}) \leq \varphi(x; \{u\}),$$

for all $u \in$.

Thesis.

Then $c(x)$ must satisfy the following properties:

(6)
$$H(x,y(x), c(x)) \equiv 0 \quad \text{for all} \quad x \in R$$
$$H(x,y(x), c(x)) = \max_{u \in U} H(x,y(x),u)$$

that is, equivalently:

(7)
$$f(x,c(x))\cdot \text{grad}\ \varphi(x;\{c\}) = -\alpha(x,c(x))$$
$$\alpha(x,c(x)) + f(x,c(x))\cdot \text{grad}\ \varphi(x;\{c\}) \equiv \underset{u \in U}{\text{Min}}\ \alpha(x,u) + $$
$$+ f(x,u)\cdot \text{grad}\ \varphi(x;\{c\})$$

Corollary 1.

If there exists a unique $c \to \bar{c}(x,y)$ so that

$$(8) \qquad y \cdot f(x, \bar{c}) - \alpha(x, \bar{c}) = \max_{u \in U} y \cdot f(x, u) - \alpha(x, u)$$

then $c(x) = \bar{c}(x, -\text{grad}\ \varphi(x))$, and the first of equations (13) and the second of equations (14) can be combined into the single property that the nonlinear Hamiltonian-Jacobi partial differential equation

$$(9) \qquad f(x,\bar{c}(x, -\text{grad}\ \varphi)) \cdot \text{grad}\ \varphi = -\alpha(x,\bar{c})x, -\text{grad}\ \varphi))$$

should have a positive definite solution on R.

Proof: It is immediate from the Fundamental Theorem:

Corollary 2. For every $x_o \in R$ it holds:

$$(10) \qquad \dot{x} = f(x,c(x)) \equiv f(x,\bar{c}(x,y)) = \text{grad}_{(y)}H(x,y)\bar{c}(x,y))$$

with $x(0) = x_o$.

$$(11) \qquad \dot{y} = -f_x{}^*(x,\bar{c}(x,y))y + \text{grad}_{(x)}\alpha(x,\bar{c}(x,y) = -\text{grad}_{(x)}H(x,y,c(x,y))$$

$$(12) \qquad y(0) = -[\text{grad}\ \varphi(x)]_{x = x_o}$$

$$(13) \qquad \dot{y}f(x,\bar{c})(x,y)) - \alpha(x,\bar{c}(x,y)) \equiv 0, \quad t \in I_+$$

Proof. We have already seen in Lemmas 3 and 4 that (6) and (7) hold. Also (10) holds by definition of c and $\ddot{c}$. Now note that with

$$y(t) = y(x(t))$$

it holds, by Lemma 5:

$$(14) \qquad \dot{y} = -[\text{grad } \varphi(x(t))]_x f(x,\bar{c}(x,y))$$

for any $x \in R$.

On the other hand, by (6) - (7) we have (13), with $\bar{c}(x,y) = c(x)$, whence

$$(15) \qquad - \text{grad } \varphi \;_x f(x,\bar{c}(x,y) + f_x^*(x,\bar{c}(x,y))y - \text{grad}_{(x)}\alpha(x,\bar{c}(x,y))$$

$$+ c_x^* \;\; f_u^*(x,\bar{c}(x,y))y - \text{grad}_{(u)}(x,\bar{c}(x,y)) \quad = 0.$$

But by Lemma 5, the coefficient of c_x^* is zero. Hence by (14) - (15), (11) must hold.

$$(16) \qquad H(x,y,c(x)) = \bar{H}(x,y)$$

for every state x and co-state $y = -\text{ grad } \varphi(x)$. Another way of expressing (16) is to say that

$$(17) \qquad c(x) = \bar{c}(x,y), \quad y = -\text{ grad } \varphi(x).$$

This fundamental result can be summarized by the statement that "a control system is optimal if, in addition to $H = 0$, one has also $H = \overline{H}$".

A <u>sufficient</u> condition for the existence of an optimal control $c(x)$ on R is that, on R, the <u>Hamilton-Jacobi Partial Differential Equation</u>

$$(18) \qquad H(x,y,\overline{c}(x,y)) = 0, \qquad y = -\operatorname{grad}\varphi$$

has in R a positive definite solution $\varphi(x)$. A more explicit way of writing (18) is

$$(19) \qquad f(x,\overline{c}(x,-\operatorname{grad}\varphi))\cdot\operatorname{grad}\varphi = -\alpha(x,\overline{c}(x,-\operatorname{grad}\varphi)).$$

This nonlinear partial differential equation appears to be quite formidable; however, for the case of a linear plant, $f(x,u) = Ax + Ku$, $\overline{c} = \operatorname{sgn}[K^* y]$, and $\alpha = 1$, it is possible to find the general solution of (19). (This last example, with constraints $|u_i| \leq 1$ and $\alpha(0,0) \neq 0$, seems to violate the preceding hypotheses; however, we shall see below how to include this case by an artifice.)

In conclusion, it is easy to remember the salient features of this (simplified) Liapunov-Pontrjagin theory of the stabilization and optimization of control systems by remembering

$$H = 0 \iff \text{STABILITY}$$

$$H = \overline{H} = 0 \iff \text{OPTIMALITY}$$

2.3 The Pay-Off Penalty and Trade-Off Functions.

2.3.1 Time-Optimal Control

Consider the system

$$\dot{x} = f(x) + Ku \qquad |u_i| \leq 1, \tag{1}$$

and try to minimize the quadratic "cost of control" criterion

$$\varphi = \varphi(x_o) = \int_0^{T(x_o)} u(t)\cdot Qu(t)dt. \tag{2}$$

We shall show that this criterion, and the constraint $|u_i| \leq 1$, lead to a dual-mode (or"linear-saturating") control law.

The Hamiltonian is

$$H = y\cdot f(x) + u\cdot\left[K^* y\right] - u\cdot Qu. \tag{3}$$

Hence

$$\mathrm{grad}_{(u)} H = K^* y - 2Qu. \tag{4}$$

A detailed study of the situation at hand shows now that

$$u = \mathrm{sat}\left[(1/2)Q^{-1}K^* y\right] \quad , \tag{5}$$

where $e^i\cdot\mathrm{sat}\left[Z\right] \triangleq e^i\cdot Z$ and where

$$\mathrm{sat}\left[\Theta\right] = \begin{cases} \Theta, & |\Theta| \leq 1; \\ \mathrm{sgn}\ \Theta, & |\Theta| \geq 1. \end{cases} \tag{6}$$

Now let $f(x,u) = Ax + Ku$. Referring to equations (18)-(23) of Example 1 it is easy to see that setting $y = - Bx$ in (5) causes (5) and the equations (22)-(23) of the following example to agree for all $\|x\| < \epsilon$, where $\epsilon < 2/\|Q^{-1}\|$, $\|K\|$ $\|B\|$. Hence we have the following result:

Let the matrix $-A$ be a stability matrix, so that the matrix equation

$$PA^* + AP = KQ^{-1}K^* \tag{7}$$

has a positive definite solution

$$P = P^* > 0. \tag{8}$$

Then the system

$$\dot{x} = Ax + Ku, \qquad |u_i| \leq 1, \tag{9}$$

becomes optimal relative to the performance criterion

$$\varphi(x_o) = \int_o^{+\infty} u(t) \cdot Qu(t)dt \tag{10}$$

if for the optimal control law $u = c(x)$ one chooses

$$c(x) = \text{sat}[Gx], \tag{11}$$

$$G = -(\tfrac{1}{2})Q^{-1}K^*B, \qquad B = P^{-1}. \tag{12}$$

EXAMPLE

Let $U = E^n$, $R = E^n$, and consider a linear plant

(1) $$\dot{x} = f(x, u) = Ax + Ku$$

with a quadratic performance index

(2) $$\alpha(x, u) = x \cdot Cx + u \cdot Qu,$$

where the matrices C and Q are positive definite; i.e., $C = C^* > 0$, $Q = Q^* > 0$. Then we have for the system's Hamiltonian

(3) $$H = H(x,y,u) = y \cdot f(x,u) - \alpha(x,u) = x \cdot A^*y + u \cdot K^*y = x \cdot Cx - u \cdot Qu$$

$$= x \cdot A^*y - x \cdot Cx + u \cdot K^*y - u \cdot Qu.$$

Now $\text{grad}_{(u)}H(x,y,\bar{c}) = K^*y - 2Q\bar{c} = 0$ whence $\bar{c} = \bar{c}(x,y)$ is given by

(4) $$K^*y = 2Q\bar{c}, \quad \bar{c} = \tfrac{1}{2} Q^{-1}K^*y.$$

Also, it is easy to verify that the Hessian at F is negative definite, i.e.,

(5) $$[\text{Grad}_{(u)}H(x,y,u)]_u = -2Q < 0,$$

whence

(6) $$\bar{c}(x,y) = \tfrac{1}{2}Q^{-1}K^*y$$

provides a true maximum to $H(x,y,u)$ on $U = E^n$. Thus we have

(7) $$\bar{H}(x,y) = H(x,y,\bar{c}(x,y)) = x\cdot A^*y - x\cdot Cx + \tfrac{1}{2}K^*y\cdot Q^{-1}K^*y - \tfrac{1}{2}K^*y\cdot Q^{-1}K^*y$$

and so the Hamilton-Jacobi equation is

(8a) $$\bar{H} = x\cdot A^*y - x\cdot Cx + \tfrac{1}{2}y\cdot (KQ^{-1}K^*)y = 0$$

(8b) $$y = -\text{grad } \varphi.$$

Also, the Hamiltonian equations corresponding to (3) and (6) are

(9a) $$\dot{x} = Ax + \tfrac{1}{2}(KQ^{-1}K^*)y, \qquad x(0) = x_o$$

(9b) $$\dot{y} = -A^*y + 2Cx, \qquad y(0) = -\text{grad } \varphi(x_o)$$

where

(10) $$\varphi(x_o) = \int_o^{+oo} [x\cdot Cx + \tfrac{1}{2}y\cdot (KQ^{-1}K^*)y]dt.$$

Now define the $2n \times 2n$ matrix $\mathcal{H}$ by

(11) $$\mathcal{H} \leq \begin{pmatrix} A & \tfrac{1}{2}KQ^{-1}K^* \\ 2C & -A^* \end{pmatrix} ,$$

and let its matrizant be given by

$$(12) \qquad \begin{matrix} M(t), & N(t) \\ R(t), & S(t) \end{matrix} \quad \triangleq e^{\quad t}$$

Then the general solution of ([illegible]) is given by

$$(13a) \qquad x(t) = M(t)x_o + N(t)y_o$$

$$(13b) \qquad y(t) = R(t)x_o + S(t)y_o \,.$$

Now suppose there exists a positive definite matrix $B = B^* > 0$ which satisfies the $\frac{1}{2}n(n + 1)$ simultaneous quadratic equations in the $\frac{1}{2}n(n + 1)$ unknowns B_{ij}, $i < j$, given by

$$(14) \qquad BA + A^*B - B(KQ^{-1}K^*)B = -C.$$

Then set

$$(15) \qquad y_o = -2Bx_o$$

and, inserting (13) into (10) and rearranging the algebra, find that $\varphi(x_o) = x_o \cdot B \cdot x_o$. However, x_o is quite arbitrary, and B is independent of x_o. Hence, in general,

$$(16) \qquad \varphi(x) = x \cdot Bx; \quad \text{grad } \varphi = 2Bx.$$

Thus equation (15) is equivalent to the statement that

(17) $$y_o = -\text{grad}\, \varphi(x_o),$$

i.e., $y(x)$ is the co-state of x. Thus we have exhibited a global solution both of the two-point boundary value problem (9) and of the Hamilton-Jacobi equation (8). Hence $c(x) = \bar{c}(x,y) = -(\frac{1}{2})[Q^{-1}K^*B]x$. Thus we have proved the following result: The system

(18) $$\dot{x} = Ax + Ku, \qquad x(0) = x_o,$$

is optimal relative to minimization of the performance criterion

(19) $$\varphi(x_o) = \int_o^{+\infty}[x(t)\cdot Cx(t) + u(t)\cdot Qu(t)]dt,$$

if the quadratic equation

(20) $$BA + A^*B - B(KQ^{-1}K^*) = -C$$

(where $C = C^* > 0$, $Q = Q^* > 0$) has a solution

(21) $$B = B^* > 0$$

and one sets

(22) $$u = Gx,$$

(23) $$G = -(\tfrac{1}{2})Q^{-1}K^*B.$$

2.3.2 Constraints.

Consider the problem of attitude control of a space vehicle by means of reaction jets. As we have seen, for time-optimal control each jet-actuated control torque (in dimensionless units) should either be given the signum +1 or -1. Such a control is called BANG-BANG control.

We shall now prove that for fuel-mass-minimal control, i.e. for the performance criterion φ defined below, the optimal attitude control must be what we shall call BANG-COAST-BANG control.

In fact, the relevant two-point boundary value problem, for the performance criterion

$$(1) \qquad \varphi = \int_0^T (1 + \mu|c|)dt, \quad \mu > 0, \qquad |c| = c \cdot \mathrm{sgn}[c] = |c_1| + \ldots + |c_n|$$

has, with a constant matrix K, the form

$$(2) \qquad \dot{x} = f(x) + Kc, \qquad x(0) = x_o,$$

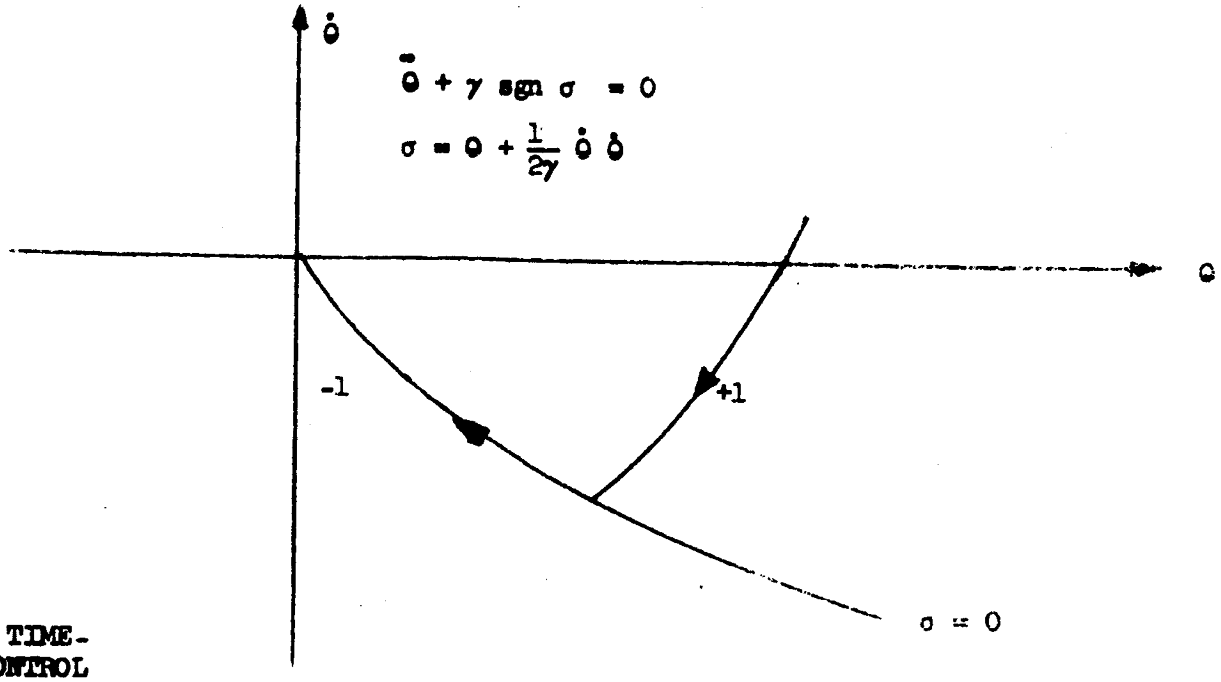
$$(3) \qquad \dot{y} = -f_x^*(x)y, \qquad y(0) = y_o,$$

$$(4) \qquad x(T) = 0, \qquad |c_i| \leq 1,$$

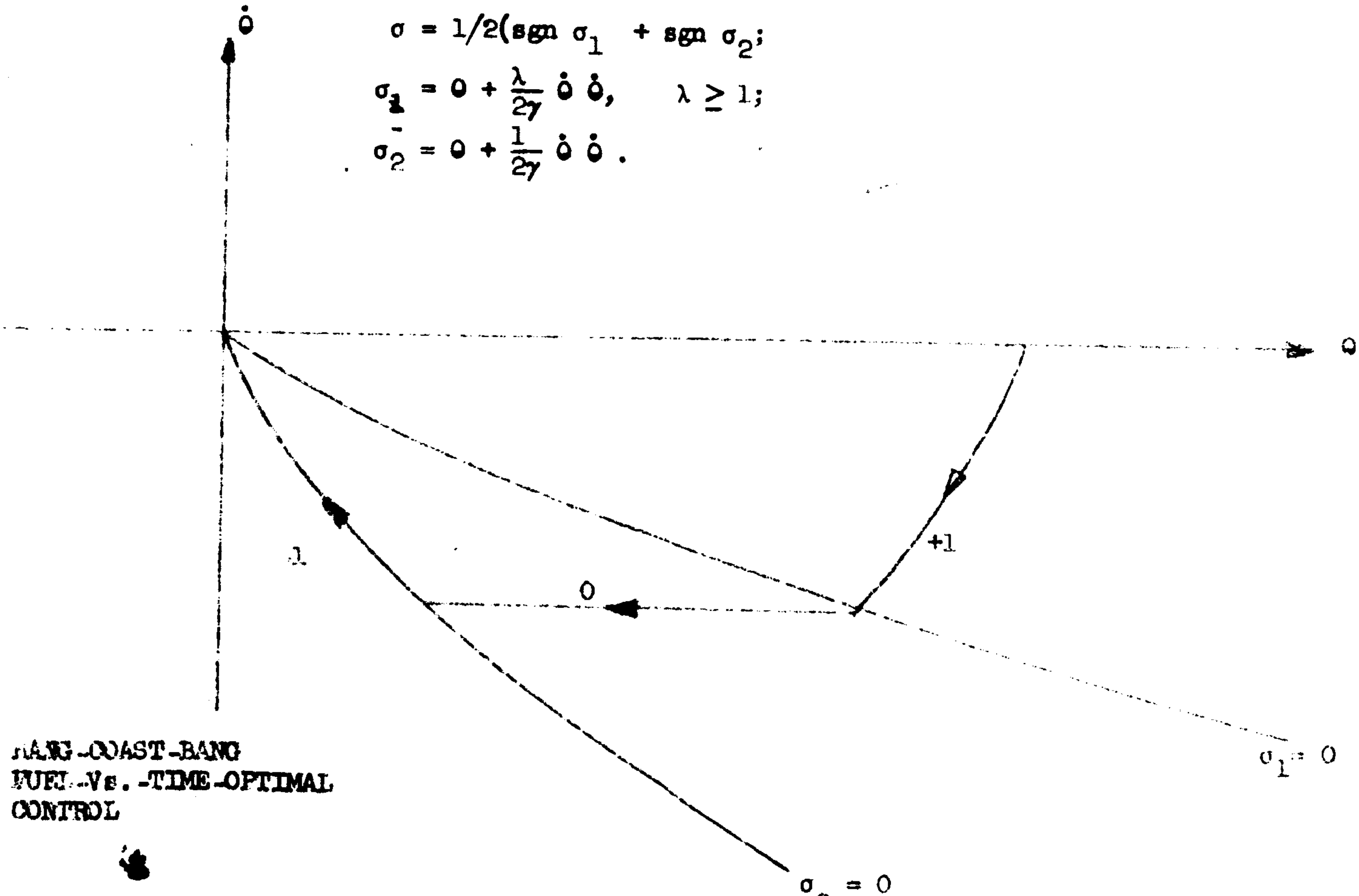
while the associated Hamiltonian is, for $\mu > 0$,

$$(5) \qquad H = y \cdot f(x) - 1 + \mu\left[c \cdot (\mu^{-1}K^* y - \mathrm{sgn}[c])\right] .$$

By the Maximum Principle $(H = \bar{H})$ we must choose c so as to maximize H relative to (4), keeping x and y fixed. This leads to the results


$\dot{\theta}$
$\ddot{\theta} + \gamma \operatorname{sgn} \sigma = 0$
$\sigma = \theta + \frac{1}{2\gamma} \dot{\theta} \dot{\theta}$
θ
-1
+1
$\sigma = 0$

BANG-BANG TIME-
OPTIMAL CONTROL


$\dot{\theta}$
$\sigma = 1/2(\operatorname{sgn} \sigma_1 + \operatorname{sgn} \sigma_2;$
$\sigma_1 = \theta + \frac{\lambda}{2\gamma} \dot{\theta} \dot{\theta}, \quad \lambda \geq 1;$
$\sigma_2 = \theta + \frac{1}{2\gamma} \dot{\theta} \dot{\theta} .$
θ
+1
1
0
$\sigma_1 = 0$
$\sigma_2 = 0$

BANG-COAST-BANG
FUEL-Vs.-TIME-OPTIMAL
CONTROL

(6) $$c_i = 1/2(1 + \mathrm{sgn}\left[|e^i\cdot K^*y| - \mu\right])\mathrm{sgn}\left[e^i\cdot K^*y\right], \qquad (i=1,\dots,n)$$

where $(e^1, e^2, \dots, e^n) = I_n$.

It is clear from the Bang-Coast-Bang control law (6) that as the fuel minimization increases in importance relative to the time-minimization, i.e., as μ increases, the law (6) gives $c_i = 0$ for longer time-intervals. (In fact, $c_i = 0$ for $t_1 < t < t_2$, whenever $|e^i\cdot K^*y(t)| < \mu$ for $t_1 < t < t_2$).

A concrete example of the principle (6) will now be given. Consider the problem of single-axis attitude control. The governing moment equation is simply

(7) $$J\ddot{\theta} + \gamma\mathrm{sgn}[\sigma] = 0.$$

It can be proved, by elementary reasoning, that for controlling (7) so that, at a pre-specified (non-minimal) time $T > 0$, one has

$$\theta(T) = \dot{\theta}(T) = 0$$

while at the same time minimizing the fuel-mass expelled

(8) $$\varphi = \int_0^T |\mathrm{sgn}[\sigma]|\,dt$$

(where now we allow $\mathrm{sgn}[\sigma] = 0$ if $\sigma = 0$), one must use the control law

(9a) $$\sigma = 1/2(\mathrm{sgn}[\sigma_1] + \mathrm{sgn}[\sigma_2]),$$

(9b) $$\sigma_1 = \theta + \lambda\frac{J}{2\gamma}\dot{\theta}|\dot{\theta}|, \quad \lambda \geq 1,$$

(9c) $$\sigma_2 = \theta + \frac{J}{2\gamma}\dot{\theta}|\dot{\theta}|.$$

It is clear from Figure 2.4.2-1 that as λ — 1, the system (7) - (9) tends to a purely time-optimal system, while $\lambda \rightarrow +\infty$, the region of coasting (with $J\ddot{\theta} = 0$, $\dot{\theta}(t) = \text{const.}$) increases and the system becomes more and more fuel-mass minimal.

The result (9) is basic to the subject of fuel-optimal attitude control.

The extension of this result to simultaneous 3-axis fuel-minimal attitude control would be extremely useful. An appropriate version of this hypothetical extension will be derived below.

Before exhibiting this 3-axis control, consider the subject of external torque disturbances. If the system (7) is subjected to a disturbing torque d, $|d| < \gamma/2$, and if d be regarded as a constant, then the time-optimal control of the system

$$(10) \qquad J\ddot{\theta} + \gamma \operatorname{sgn}[\sigma] = d,$$

is given by

$$(11) \qquad \sigma = \theta + \frac{J}{2\gamma} \cdot \frac{\dot{\theta}|\dot{\theta}|}{(1 - [d/\gamma] \operatorname{sgn} \dot{\theta}\ )} \quad .$$

In order to mechanize (11), one must measure d. This can be done by using, in (11)

$$(12) \qquad d = \gamma \operatorname{sgn}[\sigma] + \overline{J \frac{d}{dt} [\dot{\theta}]}$$

where $\operatorname{sgn}[\sigma]$ and θ are readily measurable variables. Of course, $d[\dot{\theta}]/dt$ is corrupted by noise, but filtering and smoothing techniques can be employed. By the notation

$$(13) \qquad \overline{d[\dot{\theta}]/\ dt}$$

we mean a suitably averaged and smoothed measure of $d[\dot{\theta}]/dt$.

Note that, if (12) be inserted into (11), then the system (10) is not only time-optimal but also SELF-ADAPTIVE to external disturbance torque variations. Furthermore if the values of J and γ used in (11) - (12) are not correct, we can lump the residual (or discrepancy) with d, i.e., replace d in the concept by

$$(14) \qquad d + (J - \hat{J})\ddot{\theta} + (\gamma - \hat{\gamma}) \operatorname{sgn}[\sigma]$$

where $\hat{J}$ and $\hat{\gamma}$ are the true values and J, γ the assumed values. Thus, if $(J - \hat{J})$ and $(\gamma - \hat{\gamma})$ be sufficiently small, the self-adaptive feature (11) - (12) can compensate not only for unknown external torques, (the unidentified environment) but for lack of precise identification of the system's internal characteristics. Thus the system (10) - (11) - (12) is truly self-adaptive.

In conclusion, we shall generalize this self-adaptive time-optimal control law to all 3 axes. (The extension of the fuel-minimal law is quite similar).

The system is governed by

$$(15) \qquad \dot{u}^i = w \otimes u^i, \qquad (i=1,2,3)$$

$$(16) \qquad J\dot{w} + w \otimes Jw = - \Gamma \operatorname{sgn}[g] + d$$

where $\Gamma = \operatorname{diag}(\gamma_1, \gamma_2, \gamma_2)$ and where $e^i \cdot g = \sigma_i$, $(i=1,2,3)$. Also

$$(17a) \qquad \theta_1 = -\operatorname{Arcsin}\left[u^3_2 / \sqrt{1 - (u^3_1)^2}\right]$$

$$(17b) \qquad \theta_2 = \operatorname{Arcsin}\left[u^3_1\right]$$

$$(17c) \qquad \theta_3 = -\operatorname{Arcsin}\left[u^2_1 / \sqrt{1 - (u^3_1)^2}\right] .$$

(18a) $$\dot{\Theta}_1 = w_1 + \left[\sin\Theta_1 \tan\Theta_2\right] w_2 - \left[\cos\Theta_1 \tan\Theta_2\right] w_3$$

(18b) $$\dot{\Theta}_2 = \left[\cos\Theta_1\right] w_2 + \left[\sin\Theta_1\right] w_3$$

(18c) $$\dot{\Theta}_3 = - \left[\frac{\sin\Theta}{\cos\Theta_2}\right] w_2 + \left[\frac{\cos\Theta}{\cos\Theta_2}\right] w_3 \quad .$$

Now try the control law

(19) $$\sigma_i = \Theta_i + \frac{J_i}{2\gamma_i} \cdot \frac{w_i \, |w_i|}{(1 - \frac{d_i}{\gamma_i} \operatorname{sgn}[w_i])} \quad , \qquad (i=1,2,3).$$

Choose the Liapunov function

(20) $$\varphi = \sum_{i=1}^{3} 1/2 \, J_i (w_i)^2 + \sum_{i=1}^{3} \gamma_i |\sigma_i| \, .$$

Clearly $\varphi = 0$ if and only if $\Theta_i = w_i = 0$, $(i=1,2,3)$. It can be proved that

(21) $$\dot{\varphi} \leq - (1/6) \sum_{i=1}^{3} \gamma_i |w_i| + O(\Theta_1^2, w_1^2)$$

whenever

(22) $$|d_i| < \gamma_i / 3.$$

In fact, differentiate φ with respect to time. Clearly

(23) $$\dot{\varphi} = \sum_{i=1}^{3} w_i (J\dot{w}_i) + \sum_{i=1}^{3} \gamma_i \operatorname{sgn} \sigma_i [\dot{\sigma}_i] \quad .$$

Now substitute $J\dot{w}_i$ from (16), and compute $\dot{\sigma}_i$ from (19), wherein one can substitute $\dot{\Theta}_i$ from (18) and $\dot{w}_i$ again from (16). Note that $\dot{\Theta}_i = w_i + \bigcirc(\Theta_i^2, w_i^2)$. Then by use of (22) and simple inequality arguments, the result (21) can be obtained.

If $\|u\|^2 + \|w\|^2$, or equivalently if φ be sufficiently small, then

$$\dot{\varphi} \leq -(1/12) \sum_{i=1}^{3} \gamma_i |w_i|.$$

Clearly, $\varphi(t)$ is monotone non-increasing and tends to a limit. Now $w_i \equiv 0$, ($i=1,2,3$), implies that $\dot{\varphi} = 0$, whence $\varphi = \text{const.} > 0$, whence at least one $|\sigma_i| > 0$, and so by (16), $w_i \equiv 0$ is impossible. Thus $\varphi(t) \to 0$ as $t \to +\infty$.

Therefore the control law (19) is a stable simultaneous 3-axis quasi-optimal control law in some neighborhood of the origin.

3.0 Finding the Time Optimal Switching Surfaces.*

Recently first computations of higher order switching surfaces have begun at Aeronca. Based on theoretical developments of D.C, Lewis and P. Mendelson it has been possible to carry out the computation of an explicit closed form expression for the switching surface of a plant of any order. That this has actually been done for any plant of order greater that two is in itself a milestone. Care is necessary, however, in seeing these results, now carried out only for plants of third and forth order, in proper perspective.

To date it is clear only that linear plants with a single actuator, and then only a special subclass of these, can in principle be solved. It appears, however, that the full class of linear plants with a single actuator can be successfully tackled. It is possible that extensions to many actuators and finally to nonlinear plants can, in principle, be pursued through generalizations and extensions of the methods that have brought this first problem to bay. This is how it seems at this writing. But at this writing a clear program leading to a reasonable optimal synthesis of the Saturn control seems feasible. This is because the Saturn can be reasonably approximated by a linear plant with a single actuator.

In any case, examples of optimal control laws are now in hand.

With these results, it is also apparent, that the situation is not nearly so sanguine as might have been expected from earlier confident predictions, nor of course so bleak as some investigators have painted after devoting great effort only to produce failure. After all, there are now some results in hand. The picture is actually like this. The results, that is an expression for a control surface can be found. With this expression come a set of inequalities (n-1 for an n cordinate system). These expressions, in the special case solved, besides going up linearly in number with

*This is a report of work carried out on NASA Contract NAS8-5002 the order, apparently become much more cumbersome in form. But, on the other hand, and this is of crucial importance, the procedures for obtaining the switching surfaces are very methodical despite the abstractness of the arguments.

What this means then, is, that if it is wished to be able to study cases of considerable complexity, it will be necessary to produce the expressions by a mechanical procedure. This is possible. The form of the algebraic manipulations that must be done is clear. P. Merryman has recently begun, in conjunction with this project, to reduce the algebra to machine manipulation, with a project to evaluate an arbitrary determinant whose coefficients are algebraic expressions as an algebraic expression.* A program to do this has been written and is now being readied for machine testing. This initial program was for testing the feasibility of principles that would be pertinent to such an undertaking. It is now presumed that these principles work.

Next comes the question of the applicability of the control surfaces expressions once they have been found. That guidance and insight will follow is obvious. In the context of the synthesis problem however, there are two possibilities that must be entertained. One is that the expressions generated are far too cumbersome to, with any foreseeable extrapolation of the state-of-the-art in hardware, manifest as a missile computer. In this case, these best switching surfaces would be used as a starting point for generating approximate surfaces amenable to rapid computation.

Second in this line of thought is the possibility that these are the only solutions achievable. That it will require a sophisticated perturbation theory to obtain the solutions of plants evincing even slightly more complex forms. In that

* P. Merryman has also found and programmed a procedure for finding eigenvalues and vectors of arbitrary non-singular matrices, which will be of importance in the investigation of the arbitrary linear plant. This work is reported below where the eventual application is to finding approximate optimal trajectories of general plants.

case these solutions have a central role to play, as the harmonic oscillator or the hydrogen atom in quantum mechanics. Far from being a hopeless situation, it is one in which investigators can apply a whole range of procedures to obtain the more crucial insight that they need.

From this perspective a glimmering of the programs that will be important to bring this effort to fruition are becoming clear. First to find "first integrals" from which the surfaces are built up is easy in the case attempted. In general it may be most difficult to practice the given procedure. Other procedures may have to be investigated. P. Mendelson already has some results in this direction. B. Bass also reports results of a more general nature.

Second there is the whole program of moving toward more difficult plants. As the regularities to be found have yet to be reported, it can only be commented that the construction of the surfaces should be tried in a search for these regularities. Auxilliary techniques must be found. Here, one auxilliary technique is being persued by P. Merryman. Applied to the simplification of the plant it is a procedure for finding a transformation to a simpler form. Applied to investigation for insight into the problem it is a procedure for finding approximate trajectories.

Third it will be necessary to learn how to generate mechanically the expressions for the switching surfaces as the cases are understood. To date, P. Merryman has initiated a program in this direction. P. Mendelson has hopes of working closely with him.

Fourth, a program to learn how to perturb the known solutions may have to be initiated.

Fifth, it is certain now that much will have to be learned about how to approximate the known solutions where fast computations with light equipment is imperative.

Below is presented the fundamental approach used by D.C. Lewis and P. Mendelson.

3.1 Statement of the General Problem.

We have given a system of differential equations of the form

$$(1) \qquad \dot{x} = f(x) + a\epsilon,$$

where the dot stands for differentiation with respect to the time t, where the unknown x is an n-vector, where f is an n-vector function of x, and where a is a constant non-zero n-vector. $f(x)$ is assumed to be of class C', at least. As for the scalar ϵ, this is a bounded not necessarily continuous function which is to be chosen in such a way that a solution starting with given initial conditions will be steered as quickly as possible to the origin $x = 0$. Evidently ϵ can be regarded as a function of x.

Also without essential loss of generality we can take the bound for $|\epsilon|$ to be 1. Otherwise we would modify the vector a by dividing all of its components by the bound.

From the "bang-bang" principle it is known that time optimality may be achieved in a wide variety of cases by limiting ϵ to its two extreme values $+1$ and -1. Thus, we can regard (1) as representing two systems of continuous differential equations, namely,

$$(1a) \qquad \dot{x} = F(x), \qquad \text{where} \quad F(x) = f(x) + a,$$

corresponding to $\epsilon = +1$, and

$$(1b) \qquad \dot{x} = G(x), \qquad \text{where} \quad G(x) = f(x) - a,$$

corresponding to $\epsilon = -1$. We now formulate the problem by asking how it is possible to steer a point x into the origin as quickly as possible by making it move first along a solution of the system (1a) (or (1b)) and then along a solution of (1b) (or

(1a)), and then, again, along a solution of (1a) (or (1b)), and so forth, until the origin is reached in a minimum time. The problem is to determine at what points, x, we should switch from system (1a) to (1b), or vice versa. These points are known as switching points; and point sets consisting of switching points (corresponding to all time optimal paths) are known as switching manifolds, even though these point sets need not be closed manifolds in the strict technical sense, whereby each point of the set has a neighborhood whose intersection with the set is homeomorphic to a simplex of some dimensionality ≥ 1 and $< n$. In fact most of the switching manifolds, or at least the parts of them referred to later as "leaves", will turn out to have certain boundary points which will constitute switching manifolds of lower dimensionality. Broadly speaking, our problem is to determine equations for these switching manifolds and to develop certain inequalities which can also be satisfied by points lying on the switching manifolds. These inequalities are necessary because the switching manifolds are found not to be completely determined by the equations. This is connected with the fact just mentioned that the switching manifolds are not closed.

3.2 Comments on Linear Plants.

Consider the so-called case of a controllable linear plant, whereby $f(x) = Ax$, A being an $n \times n$ constant matrix, and where the $n \times n$ matrix D, whose columns are the vectors, a, Aa, Aa^2, ..., $A^{n-1}a$, is non-singular. This definition of the controllability of a linear plant was introduced by Kalman and is designed to insure that every point in some neighborhood of the origin can be steered into the origin in the indicated manner. From this fact it is obvious that controllability is invariant under non-singular linear transformations of the vector x. Indeed it is easy to verify that if x is replaced by Lx, L being an $n \times n$ non-singular constant matrix, A must be replaced by LAL^{-1}, a by La and D by LD. And, of course, LD is non-singular, if both L and D are.

These facts make it possible to perform a preliminary normalization, so that the components of a may be assigned any special values not all zero. For instance, there is no loss of generality in assuming that the i^{th} component of A is δ_{i1}.

We next turn to a more far reaching reduction of the form of a controllable linear plant. We introduce a new unknown vector $y = D^{-1}x$, whose n-components it will be convenient to denote by $y_0, y_1, \ldots, y_{n-1}$ (rather than by $y_1, y_2, \ldots, y_n$). Then evidently $x = Dy$ and, from the original equations of the linear plant, which we recall are

$$\dot{x} = Ax + a\epsilon, \tag{2}$$

we find that

$$\dot{y} = D^{-1}\dot{x} = D^{-1}(Ax + a\epsilon) = D^{-1}ADy + D^{-1}a\epsilon.$$

Suppose that the characteristic polynomial of A is $\lambda^n - \sum_{k=0}^{n-1} p_k\lambda^k$. Note also that $x = Dy = \sum_{k=0}^{n-1} A^k a y_k$ by definition of D. Hence $ADy = \sum_{k=0}^{n-1} A^{k+1} a y_k = A^n a y_{n-1} + \sum_{k=0}^{n-2} A^{k+1} a y_k$. Hence $ADy = \sum_{k=0}^{n-1} A^{k+1} a y_k = A^n a y_{n-1} + \sum_{k=0}^{n-2} A^{k+1} a y_k$.

By the Cayley-Hamilton theorem $A^n = \sum_{k=0}^{n-1} p_k A^k$. Hence

$$ADy = \left(\sum_{k=0}^{n-1} p_k A^k\right) a y_{n-1} + \sum_{\ell=1}^{n-1} A^\ell a y_{\ell-1}.$$

Therefore $$D\dot{y} = ADy + a\epsilon = D\begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_{n-1} \end{pmatrix} y_{n-1} + D\begin{pmatrix} 0 \\ y_0 \\ y_1 \\ \vdots \\ y_{n-2} \end{pmatrix} + a\epsilon.$$

Multiplying by D^{-1}, we thus get the following equations for the linear plant when expressed in terms of $y_o, \dots, y_{n-1}$.

$$
(3) \quad \begin{cases}
\dot{y}_o = p_o y_{n-1} + 0 + b_1 \epsilon \\
\dot{y}_1 = p_1 y_{n-1} + y_o + b_2 \epsilon \\
\quad \vdots \\
\dot{y}_k = p_k y_{n-1} + y_{k-1} + b_{k+1} \epsilon, \qquad k = 1,2,\dots,n-1.
\end{cases}
$$

Here we use $b_1, b_2, \dots, b_n$ to represent the components of the n-vector $b = D^{-1}a$. This means that Db = a, so that

$$
(4) \quad \begin{pmatrix}
a_1 & \sum\limits_i A_{1i} a_i & \sum\limits_i A^{(2)}_{1i} a_i & \dots \\
a_2 & \sum\limits_i A_{2i} a_i & \sum\limits_i A^{(2)}_{2i} a_i & \dots \\
\vdots & \vdots & \vdots & \\
a_n & \sum\limits_i A_{ni} a_i & \sum\limits_i A^{(2)}_{ni} a_i & \dots
\end{pmatrix}
\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}
=
\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} ,
$$

where we have used $A^{(k)}_{ij}$ to represent the element in the i^{th} row and j^{th} column of A^k. From Cramer's rule, it is clear that (4) implies that $b_1 = 1$, while $b_2 = b_3 = \dots = b_n = 0$. Hence, from (3), we see that any controllable linear plant can be written in the prepared form

$$
(5) \quad \begin{cases}
\dot{y}_o = p_o y_{n-1} + \epsilon \\
\dot{y}_k = p_k y_{n-1} + y_{k-1}, \qquad k=1,2,\dots,n-1.
\end{cases}
$$

Notice that it is easy to eliminate $y_o, y_1, \dots, y_{n-2}$ from these equations, the result being

$$(6) \qquad y_{n-1}^{(n)} - \sum_{k=0}^{n-1} p_k y_{n-1}^{(k)} = \epsilon.$$

After (6) has once been integrated the function $y_{n-2}, y_{n-3}, \dots, y_o$ can be found successively without further integration from the last $n-1$ equations in the system (5).

This is a major conclusion: A controllable linear plant consisting of a system of n first order differential equations can always be expressed as a single n^{th} order differential equation of the form (6). The converse position is also true. For, if (6) is given a priori, we can form the system (5), which is certainly controllable, since the matrix D pertaining to (5) may be seen by a short calculation to be merely the unit $n \times n$ matrix. Of course, this means that not every system (2) is controllable. For an example we need only to choose A so that it has a pair of equal roots with simple elementary divisors.

3.3 Review of the Theory of First Integrals.

Our general method for obtaining the switching manifolds of a system such as (2) or even (1), where f need not be linear, depends upon a familiarity with the theory of first integrals. We propose here to review the simple facts needed in the following sections, where we shall explain and illustrate our method.

A first integral of the system (1a), say, is a scalar differentiable function, $\varphi(x,t)$, of the n-vector x and the scalar t, such that

$$(7) \qquad \frac{\partial\varphi}{\partial x}F(x) + \frac{\partial\varphi}{\partial t} \equiv 0.$$

This means that $\varphi[x(t),t]$ = constnat, whenever $x(t)$ is a solution of (1a). It is easy to see that (7) is necessary as well as sufficient for the constancy of any $\varphi[x(t),t]$, for the initial point $x(0)$ may be taken arbitrarily.

If $\frac{\partial\varphi}{\partial t} \equiv 0$, the first integral is said to be time-independent. Otherwise, it is called a time-dependent first integral.

Finally we will consider k-vector first integrals, both time dependent and time independent. The definition is the same as in the scalar case but in (7), φ is now interpreted as a k-vector function instead of a scalar function. Each component of a vector first intefral is, of course, a scalar first integral.

It is easy to obtain n-vector first integrals of the system (1a) by appeal to the existence theorem for such a system.

We hereby assume that F is of class C' and then we know that an n-vector differentiable function $\Psi(x_o,t)$ can be found, which, for constant n-vector x_o, is (considered as a function of t) a solution of (1a) and reduces to x_o when $t=0$. Moreover if we write $x = \Psi(x_o,t)$, we can immediately solve for x_o in terms of x and t. This is because x and x_o represent points on the same trajectory; either may be regarded as the initial point; x appears on the trajectory at time $+t$ after x_o, while x_o appears on the trajectory at time -1 after x. Hence $x_o = \Psi(x,-t)$. In other words, if $x(t)$ is any solution of the automonous system (1a), we have identically $\Psi[x(t),-t] \equiv x(0)$, which is constant. Thus $\Psi(x,-t)$ is a time dependent n-vector first integral.

It should also be stated that the corresponding n-scalar first integrals furnished by the components of $\Psi(x,-t)$ are independent in the sense that the jacobian determinant of the Ψ's with respect to the x's is never zero. It is satisfactory for our purposes to know that this is true for all small t, as it is obvious from continuity because of the fact that the Jacobian is clearly unity when $t=0$, this last fact being obvious from the identities $\varphi(x,0) = x$. To prove the statement for large t, we could mention that the Jacobian is a Wronskian of a certain set of solutions of the system of linear differential equations adjoint to the variational equations based on the solution $x = \Psi(x_0,t)$. We shall omit details on this.

What about time-independent first integrals. To discuss these, we limit all attention to a region where one of the components of the vector F does not vanish. This is apparently the case in applications to control theory. In order to single out the particular component of F which does not vanish, we change our notation. In the rest of this section, x will denote an $(n-1)$-vector, y will denote a scalar, and the system (1a) will appear in the form,

$$(8) \qquad \frac{dx}{dt} = f(x,y), \quad \frac{dy}{dt} = g(x,y), \quad g(x,y) \neq 0,$$

where, now, f is an $(n-1)$-vector continuously differentiable function of x and y. In other words the pair (x,y) replaces the previous x and the pair (f,g) replaces the previous F. Let us write the initial value solution of (8) in the form,

$$(9) \qquad x = \varphi(x_0,y_0,t), \quad y = \Psi(x_0,y_0,t),$$

where φ is an $(n-1)$-vector function and Ψ is a scalar function, and where, of course $\varphi(x_0,y_0,0) \equiv x_0$, $\Psi(x_0,y_0,0) \equiv y_0$. Then, from the previous discussion, we know that $\varphi(x,y,-t)$ constitutes an $(n-1)$-vector first integral and $\Psi(x,y,-t)$ is a scalar first integral, both of them being, in general, time-dependent.

One of the equations which expresses the fact that (9) constitutes a solution of (8) is

$$\Psi(x_0,y_0,t) = g[\varphi(x_0,y_0,t), \Psi(x_0,y_0,t)].$$

Since g is, by hypothesis, never zero in the region considered, it is clear that the derivative of Ψ with respect to t is never zero. Hence, if k is any convenient constnat (to be regarded as definitely fixed from now on), we may solve the equation

(10) $$\Psi(x,y,-t) = k$$

for t as a function of x and y, say,

(11) $$t = \tau(x,y), \quad \text{where} \quad \Psi(x,y,-\tau(x,y)) \equiv k.$$

We next define the (n-1)-vector function $\Phi(x,y)$ as follows

(12) $$\Phi(x,y) = \varphi(x,y,-\tau(x,y)).$$

The claim is now made that Φ is an (n-1)-vector time independent first integral of (8) and that $\tau(x,y)-t$ is a scalar time dependent first integral.

Proof: Since $\varphi(x,y,-t)$ and $\Psi(x,y,-t)$ are (time dependent) first integrals, we have

(13) $$\varphi_x(x,y,-t)f(x,y) + \varphi_y(x,y,-t)g(x,y) - \varphi_{-t}(x,y,-t) \equiv 0,$$

(14) $$\Psi_x f + \Psi_y g - \Psi_{-t} \equiv 0,$$

as identities in x, y and t.

Since $\Psi(x,y,-\tau(x,y)) \equiv k$, we also have

(15) $$\Psi_x - \Psi_{-t}\frac{\partial\tau}{\partial x} \equiv 0, \text{ and } \Psi_y - \Psi_{-t}\frac{\partial\tau}{\partial y} \equiv 0.$$

Therefore

(16) $$(\Psi_x - \Psi_{-t}\frac{\partial\tau}{\partial x})f + (\Psi_y - \Psi_{-t}\frac{\partial\tau}{\partial y})g \equiv 0.$$

These are identities in x and y. Subtracting (14) with t set equal to $\tau(x,y)$ from (16), we have $\Psi_{-t}(t,x,-\tau(x,v))[1 - \frac{\partial\tau}{\partial x} - \frac{\partial\tau}{\partial y}] \equiv 0$. Since Ψ_t never vanishes, we have

$$\frac{\partial\tau}{\partial x}f + \frac{\partial\tau}{\partial y}g - 1 \equiv 0 \tag{17}$$

and this expresses the fact that $\tau(x,y) - 1$ is a time dependent first integral of (8). By definition of Φ, we have

$$\frac{\partial\Phi}{\partial x}f + \frac{\partial\Phi}{\partial y}g = [\varphi_x - \varphi_{-t}\frac{\partial\tau}{\partial x}]f + [\varphi_y - \varphi_{-t}\frac{\partial\tau}{\partial y}]g.$$

From (17) this reduces to $\varphi_x f + \varphi_y g - \varphi_{-t}$, which, in turn, from (13) reduces to 0. This completes the proof of the claim.

The (n-1) components of Φ together with τ are seen to have a non-vanishing jacobian with respect to the components of x and y. In fact, from (15) and (12) we see that

$$\begin{vmatrix} \frac{\partial\Phi}{\partial x} & \frac{\partial\Phi}{\partial y} \\ \frac{\partial\tau}{\partial x} & \frac{\partial\tau}{\partial y} \end{vmatrix} = \frac{1}{\Psi_{-t}} \begin{vmatrix} \varphi_x - \varphi_{-t}\frac{\partial\tau}{\partial x} & \varphi_y - \varphi_{-t}\frac{\partial\tau}{\partial y} \\ \Psi_x & \Psi_y \end{vmatrix} .$$

Since $\frac{\partial\tau}{\partial x} = \Psi_x / \Psi_{-t}$ and $\frac{\partial\tau}{\partial y} = \Psi_y / \Psi_{-t}$ we may add to the first (n-1) rows of this determinant the last row multiplied by the (n-1)-vector φ_{-t}/Ψ_{-t}. This shows that the jacobian in question differs from the jacobian of the n time dependent first integrals, φ, Ψ, only by the non-vanishing factor Ψ_{-t}.

Thus we have proved that the transformation

$$\begin{cases} \zeta = \Phi(x,y) \\ \eta = \tau(x,y) \end{cases} \tag{18}$$

is non-singular at least in a neighborhood of any point where $g \neq 0$. With the help of this transformation, the equations (8) are reduced to the very simple form

$$(19) \qquad \frac{d\zeta}{dt} = 0, \quad \frac{d\eta}{dt} = +1.$$

The possibility of carrying out the reduction (1a) to the form (19) in a neighborhood of a point where not all components of F vanish together with a similar (but not simultaneous) reduction of (1b) is the basis of our theory of a "closed form" method of optimal control, as we shall explain in the next section. The "closed form" will involve only functions which appear in the initial value solutions of (1a) and (1b)

3.4 <u>General Method for Obtaining Switching Manifolds</u>.

We now are in a position to return to the problem previously posed with regard to the linear or nonlinear plant represented by (1) or by (1a) and (1b). In considering these systems of differential equations we consider three sets of variables as follows:

<u>The first set of variables</u> are the components of the original n-vector x, in which we have the system (1a) in the form,

$$(20) \qquad \dot{x}=F(x)$$

and the system (1b) in the form,

$$(21) \qquad \dot{x}=G(x).$$

<u>The second set of variables</u> are the components of an n-vector y, obtained by a one-to-one transformation of class C' from x in such a manner that the system (1a) appears in the simple form

$$(22) \qquad \dot{y}_i = \delta_{i1}, \qquad i+1,2,\ldots,n,$$

while the system (1b) appears in a possibly much more complicated form such as

(23) $$\dot{y} = K(y).$$

The third set of variables, components of an n-vector z, on the other hand, leave the system (1a) in a possibly very complicated form such as

(24) $$\dot{z} = L(z)$$

but have the virtue of reducing the system (1b) to the simple form,

(25) $$\dot{z}_i = \delta_{in}, \qquad i=1,2,\dots,n.$$

It is assumed that we have equations of transformation leaving the origin invariant, and valid in a neighborhood of the origin, which enable us to pass freely from any one of these three systems of variables to either of the other two. The possibility of obtaining such transformations with the desired properties is clearly indicated in the previous section, at least if $F(0) \neq 0$ and $G(0) \neq 0$, as we hereby assume.

As a point is successfully steered into the origin, it must, after its last switching, be on the half-trajectory of (1a), or of (1b), which terminates at the origin as t monotonically increases and approaches a certain terminal value T. Of course, if the point was originally on either one of these half-trajectories, it can be trivially steered into the origin with no switches whatsoever. Any other point must first be steered to one or the other of these two half-trajectories before it can reach the origin and must therefore experience a switching at some point of these half-trajectories. Moreover, this switching may occur at any point of the half-trajectories depending upon the initial position. Hence these half-trajectories constitutes a one-dimensional switching manifold R_1. It has two "leaves", $R_{1,1}$, the half-trajectory system (1a), and $R_{1,2}$, the half-trajectory of system (1b). $R_1 = R_{1,1} \cup R_{1,2}$.

For the sake of brevity, we will describe in detail only $R_{1,1}$ and the leaves of switching manifolds of higher dimensionality on whose boundary $R_{1,1}$ lies. Similar considerations may be supplied by the reader for $R_{1,2}$.

From (22) it is obvious that $R_{1,1}1$ when expressed in terms of the y's consists of those points for which $y < 0$ and $y_i = 0$, for $i=2,3,\ldots,n$. When we make a transformation to the z's, these conditions take some such form as $h_1^*(z) < 0$, $h_i^*(z) = 0$, for $i=2,3,\ldots,n$. We next write these conditions in a more suitable form, by eliminating z_1 from all but one of these n conditions; the one remaining condition is the one which expresses z_1 as a function of $z_2, \ldots, z_n$, hereafter briefly denoted by the (n-1)-vector $\bar{z}$. Assuming that this elimination can be effected, we obtain (in terms of the z's) conditions of the form,

$$(26) \qquad h_1(\bar{z}) < 0, \quad z_1 = h_2(\bar{z}), \quad h_i(\bar{z}) = 0, \qquad i=3,4,\ldots,n,$$

as both necessary and sufficient that the point $z \in R_{1,1}$.

Now any point (not initially on $R_{1,1}$) being steered successfully into the origin via $R_{1,1}$ must have been proceeding along a trajectory of (1b) just before its last switching. Hence the locus of all half-trajectories of (1b) which terminate on $R_{1,1}$ must constitute a "leaf" $R_{2,1}$ of a two-dimensional switching manifold. The detailed substantiation of this statement about $R_{2,1}$ is similar to what was stated above in substantiation of the fact that $R_{1,1}$ was part of a one-dimensional switching manifold. From (25) and (26) it is clear that a point on $R_{2,1}$ is characterized by the conditions

$$h_1(\bar{z}) < 0, \quad z_1 < h_2(\bar{z}), \quad h_i(\bar{z}) = 0, \qquad i=3,4,\ldots,n.$$

When we make a transformation to the y's, these conditions take some such form as

$\varphi_1^*(y) < 0$, $\varphi_2^*(y) < 0$, $\varphi_i^*(y) = 0$, $i=3,4,\ldots,n$. We next eliminate y_1 from all but one of these n conditions; the one remaining condition is the one which expresse y_1 as a function of $y_2,\ldots,y_n$, hereafter denoted by the $(n-1)$-vector $\bar{y}$. Assuming that this elimination can be effected, we obtain (in terms of the y's) conditions of the form,

$$(27) \qquad \varphi_1(\bar{y}) < 0, \quad \varphi_2(\bar{y}) < 0, \quad y_1 = \varphi_3(\bar{y}), \quad \varphi_1(\bar{y}) = 0, \quad i=4,\ldots,n,$$

as noth necessary and sufficient that the point $y \in R_{2,1}$.

Now any point being steered successfully into the origin via $R_{2,1}$ and $R_{1,1}$ (assuming that it did not start on $R_{2,1}$), must have been proceeding along a trajectory of (1a) just before switching onto $R_{2,1}$. Hence the locus of all half-trajectories of (1a) which terminate on $R_{2,1}$ must constitute a "leaf" $R_{3,1}$ of a three-dimensional switching manifold. From (22) and (27), it is clear that a point on $R_{3,1}$ is characterized by the conditions

$$\varphi_1(\bar{y}) < 0, \quad \varphi_2(\bar{y}) < 0, \quad y_1 < \varphi_3(\bar{y}), \quad \varphi_i(\bar{y}) = 0, \quad i=4,\ldots,n.$$

This process may be continued by induction, yielding, for any positive integer $k < n$, a "leaf" $R_{k,1}$ of a k-dimensional switching manifold. This leaf is characterized by n conditions, k of which are inequalities and $(n-k)$ of which are equalities. These latter may be expressed by equating to 0 certain time-independent first integrals of (1a), if k is odd, and of (1b), if k is even.

A main purpose of this paper is to carry this procedure out in detail for the case of the linear plant of order 4 in the special case in which all eigenvalues of the matrix A vanish. In other words, the system considered can be presented in the form (5) or (6) in the special case $p_0 = p_1 = p_2 = p_3 = 0$ $(n+4)$. This example should give a good idea of the general behavior of such systems even whem the p's are not all zero, and our results obtained from a study of this simple example should approximate the results to be obtained when the p's are small. The reason for this is roughly as follows:

Our methods are based on certain transformations between the x's, y's, and z's. These transformations depend continuously upon certain systems of first integrals of (1a) and (1b), which are written down in terms of the initial value solutions of the differential systems (1a) and (1b). Now, if these systems depend continuously on certain parameters, such as the p's, it is well known that the initial value solutions likewise depend continuously on the same parameters. Hence our results will be but slightly effected by small deviations of the p's from 0.

4.0 THE ADJOINT SYSTEM AND THE MAXIMUM PRINCIPLE.

Among the major steps in the theory of synthesis of optimal control systems was that taken first by R.W. Bass (10c. cit. page 191; 1956) with the introduction of a system of differential equations for certain "conjugate" or "adjoint" variables intimately associated with the variables of the system; see also the subsequent development of the work by Desoer. A second, not unrelated step was indicated in 1958 by Pontragin in the proclamation of the "Maximum Principle" wherein the problem receives a Hamiltonian formulation. For the convenience of the reader, a summary of this theory will now be presented.

System State. A set of numbers characterizing the dynamical system to be controlled is called the system state or position in state-space. It is assumed that these numbers can be sensed instantly and precisely; in reality, stochastic consideration, filtering and prediction theory enter at this point, but the overidealization involved in this assumption is sufficient for preliminary designs. The system state is represented by a vector x (or a "point" in n-dimensional Euclidean space E^n), and the system's evolution with time is specified by the curve $x(t)$ in the state space E^n).

System Dynamics. The evolution of $x(t)$ is assumed to be determined by the differential system

$$(1) \qquad \dot{x} = f(x,c), \qquad x(0) = x_o, \qquad (\dot{} = d/dt)$$

where x_o is the initial state, and c is the control vector.

Feedback Control Function. If the vector c depends only on the state x, i.e.

$$(2) \qquad c = c(x)$$

then we have instantaneous state feedback control.

State Acquisition Problem. A typical desideratum is that the system (1) evolves so that, at some future time $T = T(x_o) > 0$, the state attains the position

$$x(T) = 0. \tag{3}$$

If this be true for every x_o in a region R containing $x = 0$, we say that the system (1) is controllable, that R is the stability domain, and that T is the transition time.

Control Constraint. A realistic assumption is that no control law is admissible unless, for each x in R,

$$c(x) \text{ is contained in } \bar{U} \tag{4}$$

where $\bar{U}$ is a closed, bounded, convex subset of E^n. The control is saturated if $c(x)$ lies in the boundary of U for the state x.

Performance Criterion. We may assume that, for every x_o, the future path $x(t)$, $0 \leq t \leq T(x_o)$ can be accurately predicted if $c(x)$ be precisely specified. Hence for each piecewise smooth function $c(x)$ satisfying the constraint (4), we may (in principle at least) compute any predicted path criterion of the type

$$\phi = \phi(x_o) = \int_0^{T(x_o)} \alpha(x, c(x))dt \tag{5}$$

where $\alpha \geq 0$ is any desired smooth function of the instanteneous system state $x(t)$ and corresponding control $c(x(t))$.

Optimal Control. An optimal control law $c(x)$ provides an absolute minimum to ϕ (≥ 0) for every x_o in R, relative to all other admissable control functions.

If this control be unique, it is called the optimal control function.

Conjugate State. Suppose that $\phi(x)$ is a smooth function almost everywhere (a.e.) in R. Then grad ϕ exists a.e. in R, and we call the vector $y = -g(x) = -\text{grad}\ \phi(x)$ the co-state of the state x.

Hamiltonian. For every state x, co-state y, and control c, we may define the corresponding Hamiltonian

$$H = H(x,y,c) = y \cdot f(x,c) - \alpha(x,c). \tag{6}$$

Maximum Principle. We define the function

$$c = \bar{c}(x,y) \tag{7}$$

for all y in E^n, by specifying that

$$H(x,y,\bar{c}) = \underset{u \text{ in } U}{\text{Max}}\ H(x,y,u). \tag{8}$$

Extremal Control. We call the control function

$$c(x) = \bar{c}(x,-g(x)) \tag{9a}$$

$$g(x) = \text{grad}\ \phi(x) \qquad (x \text{ a.e. in } R) \tag{9b}$$

an extremal feedback control function.

Synthesis of Optimal Control Systems. It can be shown that if $c(x)$ be defined everywhere in R in such a way that (9a,b) holds, a.e., and such that $x(t)$ is continuous and its derivative $x(t)$ is continuous from the right (i.e. $\dot{x}(t) = \dot{x}(t+0)$) then the system (1) with this control law is optimal relative to the given constraints and performance criterion.

Thus, in order to design and synthesize and optimal control system, given the dynamics $f(x,c)$, constraints U, and performance index $\alpha(x,c)$, one needs to derive or compute

(a) the stability domain R;

(b) the switching function $g(x)$ (for all x in R).

There are basically three distinct ways to do this, and correspondingly three distinct types of control computers.

Firstly, suppose that there exists a $T = T(x_o) > 0$ and a $y(0) = -g(x_o)$ such that the TWO-POINT BOUNDARY-VALUE PROBLEM

$$\text{(10a)} \qquad \dot{x} = f(x,\bar{c}(x,y)), \qquad x(0) = x_o, \quad x(T) = 0,$$

$$\text{(10b)} \qquad \dot{y} = -f_x{}^*(x,\bar{c}(x,y))y + \text{grad}_{(x)}\alpha(x,\bar{c}(x,y)), \qquad y(0) = -g(x_o)$$

$$\text{(10c)} \qquad g(x_o)\cdot f(x_o,\bar{c}(x_o))) = -\alpha(x_o,\bar{c}(x_o,-g(x_o)))$$

has a solution. Then

$$\text{(11)} \qquad y\cdot f(x,\bar{c}(x,y)) = \alpha(x,\bar{c}(x,y))$$

for $0 \leq t \leq T$, and

$$\text{(12)} \qquad g(x_o) = \text{grad}\ \phi(x_o).$$

We may readily solve problems (a) and (b) simultaneously by running (10) "backwards in time" from the final state $x = 0$ with every possible co-state $y(T)$ compatible with (10c) and for every transition time $T > 0$. Such a computation produces every admissable initial co-state $y(0)$ $(= -g(x_o))$ and every initial state x_o in R in an efficient and non-redundant manner. This is the second method.

The two-point boundary-value problem can be given in a concise and elegant formulation. In fact, using (6), we may re-write equations (10a,b) as a Hamiltonian system

$$\dot{x} = \text{grad}_{(y)} H(x,y,c) \qquad x(0) = x_o, \tag{13a}$$

$$\dot{y} = -\text{grad}_{(x)} H(x,y,c), \qquad y(0) = y_o, \tag{13b}$$

$$c = \bar{c}(x,y), \qquad y_o = -\text{grad}\ \phi(x_o). \tag{13c}$$

The condition (1) then becomes the requirement that the initial value of the Hamiltonian be zero, i.e., that

$$H(x_o,\ y_o,\ \bar{c}(x_o,y_o)) = 0, \tag{14}$$

while the result (11) states that the Hamiltonian is constant along any optimal curve $x(t)$, i.e.,

$$H(x(t),\ y(t),\ \bar{c}(x(t),\ y(t))) = 0, \qquad 0 \leq t \leq T. \tag{15}$$

The third method is more difficult to use but preferable where possible. This is to find explicitly scalar function $T(x) > 0$ and $\phi(x) > 0$ which satisfy a.e. in some domain R the partial differential equations

$$f(x,c) \cdot \text{grad}\ T(x) = -1 \tag{16a}$$

$$f(x,c) \cdot \text{grad}\ \phi(x) = -\alpha(x,c) \tag{16b}$$

$$c = \bar{c}(x, -\text{grad}\ \phi(x)). \tag{16c}$$

We shall now apply the preceding theory to the problem of satellite attitude control of an orbiting vehicle.

For the attitude control system under consideration, and for an arbitrary permissible performance index, $\alpha(u^1, u^2, u^3, w)$ the first method of computing the optimal control function may be summarized as follows:

FUNDAMENTAL THEOREM. Consider the system

$$(17) \qquad x^i = -p \otimes x^i, \qquad x^i(0) = u^{i,1} \qquad (i=1,2,3)$$

$$(18) \qquad J\dot{p} - p \otimes Jp = -g \operatorname{sgn}\left[G^* q\right], \qquad (p(0) = 0),$$

$$(19) \qquad \dot{y}^i = -q \otimes y^i - \operatorname{grad}_{(u^i)} \alpha(x^1, x^2, x^3, p) \qquad y^i(0) = y_o^i , \qquad (i=1,2,3)$$

$$(20) \qquad \dot{q} = \sum_{i=1}^{3} x^i \otimes y^i - \bar{K}(p) \Lambda q - \operatorname{grad}_{(w)} \alpha(x^1, x^2, x^3, p) \qquad q(0) = q_o$$

where

$$\bar{K}(p) \Lambda q = J \left[(J^{-1} p) \otimes q \right] - J^{-1} \left[(Jp) \otimes q \right] ,$$

and

$$(21) \qquad \sum_{i=1}^{3} \| y_o^i \|^2 + \| q_o \|^2 = 1.$$

For each $t > 0$, put

$$(22) \qquad u^i = x^i(t), \qquad (i=1,2,3)$$

$$(23) \qquad w = p(t)$$

and define c by

$$(24) \qquad c = c(u^1, u^2, u^3, w) - \operatorname{sgn} G^* q(t).$$

The control c is permissible and optimal with respect to the criterion

$$\phi = \int_{o}^{t} \alpha ds.$$

Not that as the initial vectors vary over the sphere (21), the states for which c can be defined (22) - (24) fill the entire six-dimensional manifold.

$$\|u^i\| = 1, \quad (i=1,2.), \quad -\infty < w_j < \infty, \qquad (j=1,2,3).$$

Thus by integrating the system for a sufficiently dense set of initial conditions, the optimal control vector c, can be determined on an arbitrarily dense set of system states.

4.1 A General Two-Point Boundary Value Problem

The optimal control problem leads to an equivalent two-point boundary-value problem which is stated as follows.

We are given n linearly independent vectors ℓ^i, $(i=1,\ldots,n)$ and n scalars α_i and β_j, with $i=1,2,\ldots,p$, $j=1,2,\ldots,q$, and $p+q=n$. The problem is to find n-vectors x^o, x^N and a value of the independent variable (time) $T > 0$ such that the differential vector equation

$$\text{(1)} \qquad \frac{dx}{dt} = f(x) \qquad (x \text{ in } E^n)$$

has a solution $x = x(t)$ on $0 \leq t \leq T$ which satisfies

$$\begin{array}{lll} \text{(2)} \quad \text{(i)} & x(0) = x^o \text{ and } x(T) = x^N & \\ \quad\quad \text{(ii)} & \ell \cdot x^o = \alpha_i, & (i=1,2,\ldots,p) \\ \quad\quad \text{(iii)} & \ell^{p+j} \cdot x^N = \beta_j & (j = 1,2,\ldots,2 \text{ and } p+q=n) \end{array}$$

This is the classical two-point boundary value problem which has in general no closed form solution. One practical way of solving the optimal control problem consists in mechanizing an approximate solution of this problem. This can be achieved by digital real-time computation.

4.2 Finite-Differences Approximation.

We want to determine a discrete approximation of $x(t)$, the solution of the differential equation (1) of the preceding chapter with the conditions (2 i, ii, iii) of the preceding chapter.

For some large integer $N > 0$, find a $T > 0$ and a sequence $\left\{x^i\right\}$, $(j=0,1,\ldots N)$ of n-vectors which satisfies

$$(1) \qquad \ell^i \cdot x^o = \alpha_i, \quad (i=1,\ldots,p); \quad \ell^{p+j} \cdot x^N = \beta_j, \quad (j=1,\ldots,q)$$

and which minimizes the non-negative scalar error function $\rho = \rho(x^o, x^1, \ldots, x^{N-1}, x^N; T)$ defined by

$$(2) \quad 2\rho = \sum_{j=1}^{N-1} \left[\left\| x^j - x^{j-1} - \left(\frac{T}{N}\right) f(x^{j-1}) \right\|^2 + \left\| x^j - x^{j+1} + \left(\frac{T}{N}\right) f(x^{j+1}) \right\|^2 \right] +$$

$$+ \left\| x^o - x^1 + \left(\frac{T}{N}\right) f(x^1) \right\|^2 + \left\| x^N - x^{N-1} - \left(\frac{T}{N}\right) f(x^{N-1}) \right\|^2.$$

Note that if $\rho = 0$, then, approximately,

$$(3) \qquad x^j = x(t_j), \quad t_j = j(T/N), \qquad (j=0,1,\ldots,N)$$

$$x(t_j) = x(0) + \int_o^{t_j} f(x(\theta))d\theta, \qquad (j=0,1,\ldots,N)$$

4.3. Iterative Relaxation Algorithm.

Let N be fixed. Let $\{x^{j,0}\}$, $(j=0,1,\ldots,N)$ be an arbitrary sequence satisfying

$$(1) \qquad \ell^{i}\cdot x^{o,o} = \alpha_i, \quad (i=1,\ldots,p) \quad \ell^{p+j}\cdot x^{N,o} = \beta_j, \quad (j=1,\ldots,q), \quad p+q = n.$$

For $(v=0,1,2,\ldots)$ let $\{x^{j,v+1}\}$, $(j=0,1,\ldots,N)$, be sequences satisfying

$$(2) \qquad \ell^{i}\cdot x^{o,v+1} = \alpha_i, \quad (i=1,\ldots,p); \quad \ell^{p+j}\cdot x^{N,v+1} = \beta_j, \quad (j=1,\ldots,q), \quad p+q = n,$$

and defined inductively by

$$(3) \qquad x^{j,v+1} = F^{j}(x^{j-1,v}, \; x^{j,v}, \; x^{j+1,v}; \; T_v), \qquad (j=1,2,\ldots,N-1),$$

$$(4) \qquad T_{v+1} = \Phi(x^{o,v}, \; x^{1,v}, \; \ldots, \; x^{N,v}),$$

$$(5) \qquad F^{o}(x^{o,v+1},T_v) = \tilde{F}^{o}(x^{1,v},T_v), \qquad \tilde{F}^{o}(x^{N,v+1},T_v) = F^{o}(x^{N-1,v},T_v),$$

$(v=0,1,2,\ldots)$ where F^{j}, $(j=1,2,\ldots,N-1)$, Φ, F^{o}, $\tilde{F}^{o}$ are defined by

$$(6) \qquad F^{j} = F^{j}(x,y,z,T) = $$

$$= 1/2(x+z) - \frac{T}{4N}\left\{f(z) - f(x) - f^{*}_{x}(y)\,[z-x]\right\} +$$

$$- 1/2\left(\frac{T}{N}\right)^{2} f^{*}_{x}(y)f(y), \qquad (j=1,2,\ldots,N-1),$$

$$(7) \qquad \Phi = \Phi\,(x^{o}, \, x^{1}, \, \ldots, \, x^{N}) = \frac{N\Phi_o}{\Phi_1} =$$

$$= \frac{N\left\{\sum_{j=1}^{N-1}\left[(x^{j} - x^{j-1})\cdot f(x^{j-1}) + (x^{j+1} - x^{j})\cdot f(x^{j+1})\right]\right\}}{\sum_{j=1}^{N-1}\left[\|f(x^{j-1})\|^{2} + \|f(x^{j+1})\|^{2}\right]},$$

(8) $$F^{\circ}(x,T) = x + \left(\frac{T}{2N}\right) f(x) \ ,$$

(9) $$\tilde{F}^{\circ}(x,T) = x - \left(\frac{T}{2N}\right) f(x)$$

4.4. Fundamental Relaxation Theorem.

THEOREM. If the sequences $\left\{x^{j,v}\right\}$, $(j=0,1,\dots,N)$ *converge**, i.e. *if there exist*

(1) $$x^{j} = \lim_{v \to +\infty} x^{j,v} ,$$

then $\left\{x^{j}\right\}$, $(j=0,1,\dots,N)$ *is a solution of the finite differences approximation to the two-point boundary-value problem.*

PROOF. Note that for $j=1,2,\dots,N-1$,

(2) $$2\rho = \dots + \left\|x^{j-1} - x^{j} + \left(\frac{T}{N}\right) f(x^{j})\right\|^{2} + \\ + \left\|x^{j} - x^{j-1} - \left(\frac{T}{N}\right) f(x^{j-1})\right\|^{2} + \left\|x^{j} - x^{j+1} + \left(\frac{T}{N}\right) f(x^{j+1})\right\|^{2} + \\ + \left\|x^{j+1} - x^{j} - \left(\frac{T}{N}\right) f(x^{j})\right\|^{2} + \dots \quad .$$

Hence

(3) $$\operatorname{grad}_{(x^{j})} \rho = 4\left[x^{j} - F^{j}(x^{j-1}, x^{j}, x^{j+1}, T)\right] \quad \text{for} \quad (j=1,2,\dots,N-1).$$

(4) $$\operatorname{grad}_{(x^{o})} \rho = 2\left[(F^{o}(x^{o},T) - \tilde{F}^{o}(x^{1},T)\right]$$

(5) $$\operatorname{grad}_{(x^{N})} \rho = 2\left[\tilde{F}^{o}(x^{N},T) - F^{o}(x^{N-1},T)\right]$$

(6) $$\frac{\partial\rho}{\rho T} = (1/N)\left[- \Phi_{o} + (T/N)\Phi_{1}\right] .$$

Therefore, ρ has an extremum when

*There exist sequences which converge. The conditions for convergence have been extensively studied by Richardson and other authors.

(7) $$x^j = F^j(x^{j-1}, x^j, x^{j+1}, T), \qquad (j=1,2,\ldots,N-1)$$

(8) $$F^o(x^o, T) = \tilde{F}^o(x^1, T), \quad \tilde{F}^o(x^N, T) = F^o(x^{N-1}, T),$$

(9) $$T = \Phi(x^o, x^1, \ldots, x^N).$$

4.5 Liapunov Stability of Control Based on the Approximate Closed Form Solution.

(1) $$\dot{u}^i = w \otimes u^i, \qquad (i=1,2,3)$$

(2) $$J\dot{w} + w \otimes Jw = - \Gamma \, \mathrm{sgn}[g] + d$$

where $\Gamma = \mathrm{diag}\,(\gamma_1, \gamma_2, \gamma_3)$ and where $e^i \cdot g = \sigma_i$ $\qquad (i=1,2,3)$

(3a) $$\dot{\Theta}_1 = w_1 + \left[\sin\Theta_1 \tan\Theta_2\right] w_2 - \left[\cos\Theta_1 \tan\Theta_2\right] w_3$$

(3b) $$\dot{\Theta}_2 = \left[\cos\Theta_1 \; w_2 + \; \sin\Theta_1\right] w_3$$

(3c) $$\dot{\Theta}_3 = - \left[\frac{\sin\Theta}{\cos\Theta_2}\right] w_2 + \left[\frac{\cos\Theta}{\cos\Theta_2}\right] w_3 .$$

Now use the control law

(4) $$\sigma_i = \Theta_i + \frac{J_i}{2\gamma_i} \cdot \frac{w_i \left|w_i\right|}{(1 - \frac{d_i}{\gamma_i} \,\mathrm{sgn}\left[w_i\right]} , \qquad (i=1,2,3)$$

Choose the Liapunov Function

(5) $$\varphi = \sum_{i=1}^{3} 1/2 \, J_i (w_i)^2 + \sum_{i=1}^{3} \gamma_i \left|\sigma_i\right| .$$

Clearly $\varphi = 0$ if and only if $\Theta_i = w_i = 0$, $(i=1,2,3)$. It can be proved that

(6) $$\dot{\varphi} \leq - (1/6) \sum_{i=1}^{3} \gamma_i \left|w_i\right| + \mathcal{O}(\Theta_i^{\,2}, w_i^{\,2})$$

whenever

$$(7) \qquad d_i < \gamma_i / 3 \ .$$

Consider the system (1) - (2), with (4). Define as in (5) and differentiate with respect to time. Clearly

$$(8) \qquad \varphi = \sum_{i=1}^{3} w_i (J\dot{w}_i) + \sum_{i=1}^{3} \gamma_i \operatorname{sgn} \sigma_i \, \dot{\sigma}_i \ .$$

Now substitute $J\dot{w}_i$ from (1) and compute $\dot{\sigma}_i$ from (4) wherein one can substitute $\dot{\theta}_i$ from (3) and $\dot{w}_i$ again from (1). Note that $\theta_i = w_i + \mathcal{O}(\theta_i{}^2, w_i{}^2)$. Then by use of (7) and simple inequality arguments, the result (6) can be obtained.

4.6 A Steepest Descent Algorithm for Application to Determining Optimal Control Trajectories.

This investigation has been directed at making it possible to investigate methodically the nature of optimal trajectories for the whole gamut of cases that arise. Many of these cases have properties of nonlinearity in the part of the differential equation describing the plant without the actuators. A numerical procedure promising reasonably rapid convergence to a discretized solution of the boundary value problem was sought. The special feature that it could be especially efficient in the investigation of a region of a solution was required. From this definition it was proposed to investigate the possibility of a steepest descent procedure. This investigation is now in progress with the development of computer programs for applying the principles arrived at to several problems.

On the assumption that it is possible to construct a function on a space of a dimension that is a multiple of the dimension of the phase space by the number of points that must be chosen for a reasonably good trajectory, (1) that this function is real valued, (2) that the gradient of this function can be analytically determined, (3) and that this function takes its minimum value 0 uniquely for a best approximation to the solution, it is possible to construct a deepest descent algorithm. In the case of the trajectory for both the system and the adjoint system. Much less computation is needed than would be involved in direct relaxation or by attempting to solve the system of linear equations that arise in such an approximation to a boundary value problem.

Somewhat as a surprise to the investigators it turned up that no such algorithm as the one being investigated had been applied in attempting to solve more classical problems. Such problems are the case of matrix theory where it is desired to "diagonalize" a matrix or the finding of roots of arbitrary polynomials. As the methods cover these cases as well the the one directly investigated, the technique is first being applied to the solving of polynomials. Here the ease of constructing examples allows a thorough investigation of the technique, which then in turn is to be applied to the finding of optimal trajectories. Thus, first the method is investigated then the method becomes the tool of investigation rather than the object of the investigation. The following is a discussion of the procedure as applied to the polynomial case. It is to be noted that the procedure makes none of the usual assumptions of real coefficients as in the polynomial case, or of the symmetry or of distinct eigenvalues as is usual in the case of the matrix. It should also be borne in mind particularly in the case of the matrix that it is important to have an iterative procedure for successful machine computation as the build up of error in the so called direct methods can be prohibitive. (Actually there can be no completely direct method for finding the eigenvalues of matrices of dimension greater than four).

The essence of the procedure is to construct a function that is positive definite at all points other than at the solutions where it takes the value zero. In the case of a matrix, the function is real-valued and on the n dimensional space of the vectors. Substituting a trial value for the starting vector $\bar{x}$ the gradient is determined. The remaining problem is to determine the optimal distance along the gradient. The solution of this problem requires finding the solutions of a pair of transcendental equations.

In the matrix case (or in the case of the polynomial - the companion matrix case). In the matrix case we have as in (1) where x is a vector A the matrix and λ is a scalar, as the condition of an eigenvector. Examining the matrix row by row we have the situation

$$\lambda x_i = a_i \cdot x \qquad (1)$$

where

(2) $$A = \begin{vmatrix} a_1 \\ \vdots \\ a_n \end{vmatrix} \qquad x = \begin{vmatrix} x \\ \vdots \\ x_n \end{vmatrix} \quad .$$

Now the condition that the λ be the same for each x_i is the condition for selecting the eigenvectors, and thus define a set of eigenvalues as in

(3) $$\lambda_i = \frac{a_i \cdot x}{x_i} \quad ,$$

$\lambda_i = \lambda_j$ if and only if

(4) $$\frac{a_i \cdot x}{x_i} = \frac{a_j \cdot x}{x_j} \quad ,$$

then gives the condition for solution or equivalently (4), if

(5) $$u_{ij} = x_j(a_i \cdot x) - x_i(a_j \cdot x) = 0.$$

The set of all these conditions is encompassed in the function

(6) $$\psi(x) = \sum_{ij} u_{ij}\bar{u}_{ij} = \sum_{ij}(x_j(a_i \cdot x) - x_i(a_j \cdot x))(\bar{x}_j(\bar{a}_i \cdot x) - \bar{x}_i(\bar{a}_j \cdot x))$$

and it is this function that is to be minimized. Note that the condition makes no special case of nonlinear roots of the matrix or symmetry of any kind as well as being independent of whether the coefficients, eigenvalues, or eigenvectors are complex.

The procedure followed to find the solution is that of steepest descents. Fig.7 gives a picture of the procedure. On the contour map is the trial solution x_1.

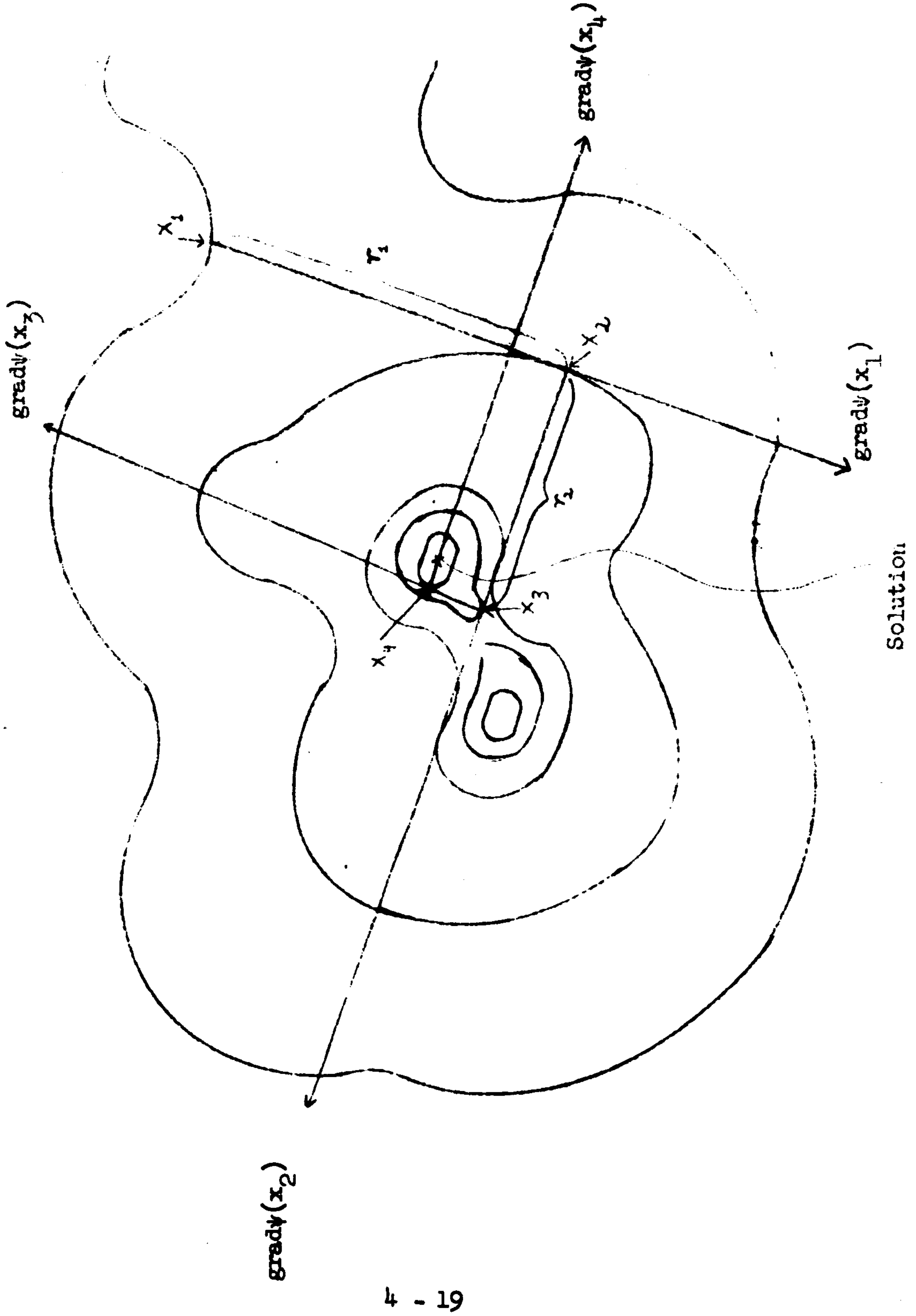


Fig. (7)

The gradient of ψ is determined at that point. Then the distance is determined r_1 in such a manner as to minimize the value that ψ can have along the line determined by the gradient. This gives the point x_2 which is used as the new starting value for the determination of the next improvement to the trial vector.

The condition for the optimum distance is embodied in

$$(8) \qquad \psi r^3 + 2r(b \cos \phi' + B) + \xi) + f \cos(\phi' + \alpha) = 0,$$

$$(9) \qquad 2r \sin(2\phi' + B) + f \sin (\phi' + \alpha) = 0.$$

These equations arise from the variation of $\psi(x + c \text{ grad } \psi(x))$ where r, ϕ' are the norm and modulus respectively of the complex number z. First the substitutions

$$(10) \qquad \phi^3 = \phi' - \frac{B}{2}$$

is made. It can then be noted that a translation of ϕ by π gives back the same solutions, (The modulus π is just - 1). Only two quadrants are pertinent to the solution. In Fig. (11) are shown the behavior of (8) and (9) over a 2 quadrant region. The x's locate the solutions, of which there can be only 7 at most. The solution 8A is either the real solution of a cubic with a single real root or its continuation in the case that there are three real roots. On the solution 8B the arrows mark the points where the solutions disappear off the real plane. A sole solution as in Fig. (12) can occur if the ratio of f to b is large enough. Finding these solutions would, of course, be tedious if this information were not known. As it is, the algorithm would be difficult if a direct attack were made. If, however, the ratio of b to f is adequately large, the diagram takes on the following for Fig. (13). In this case, the solutions can be written down approximately by

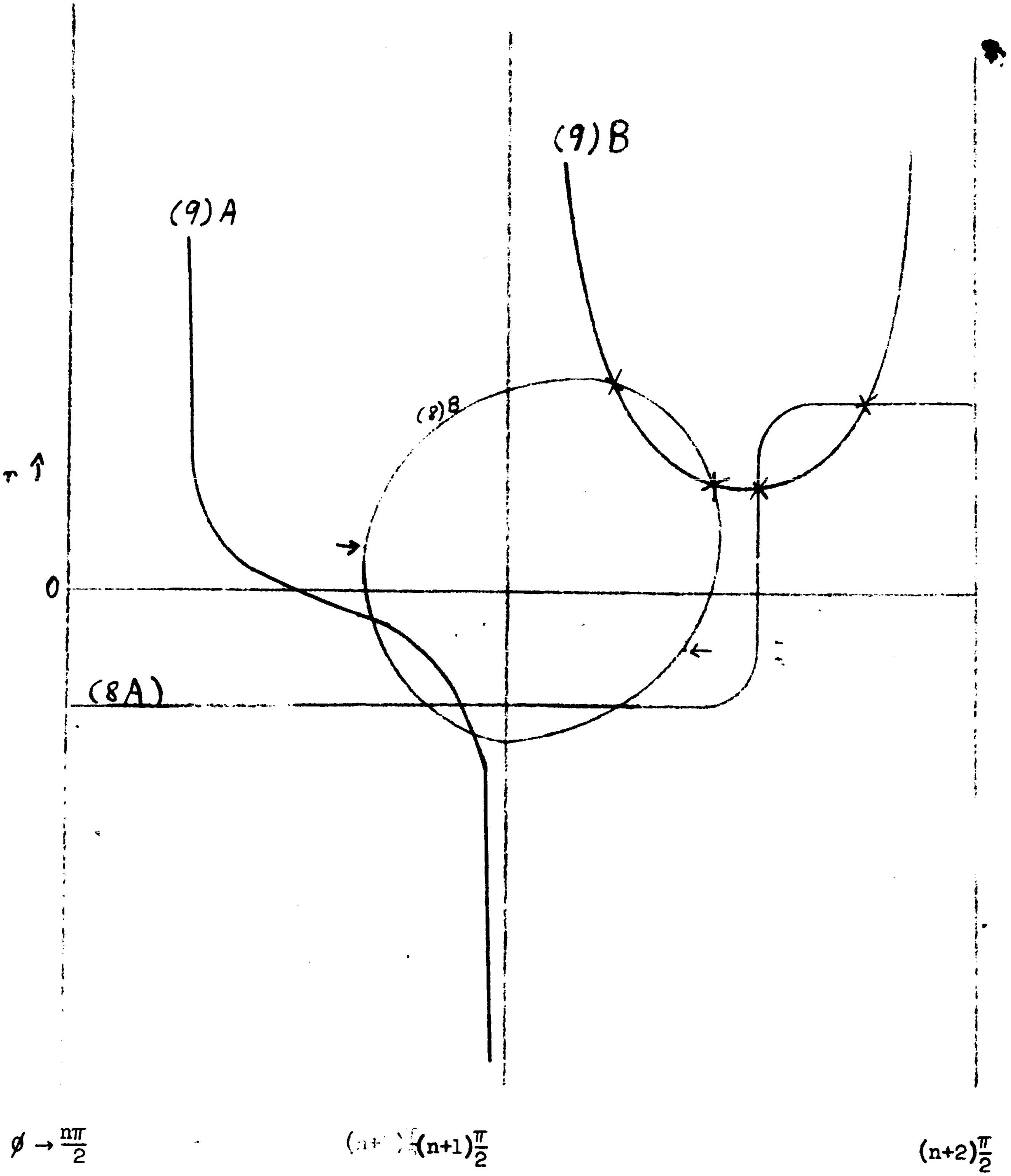


Fig. (11)

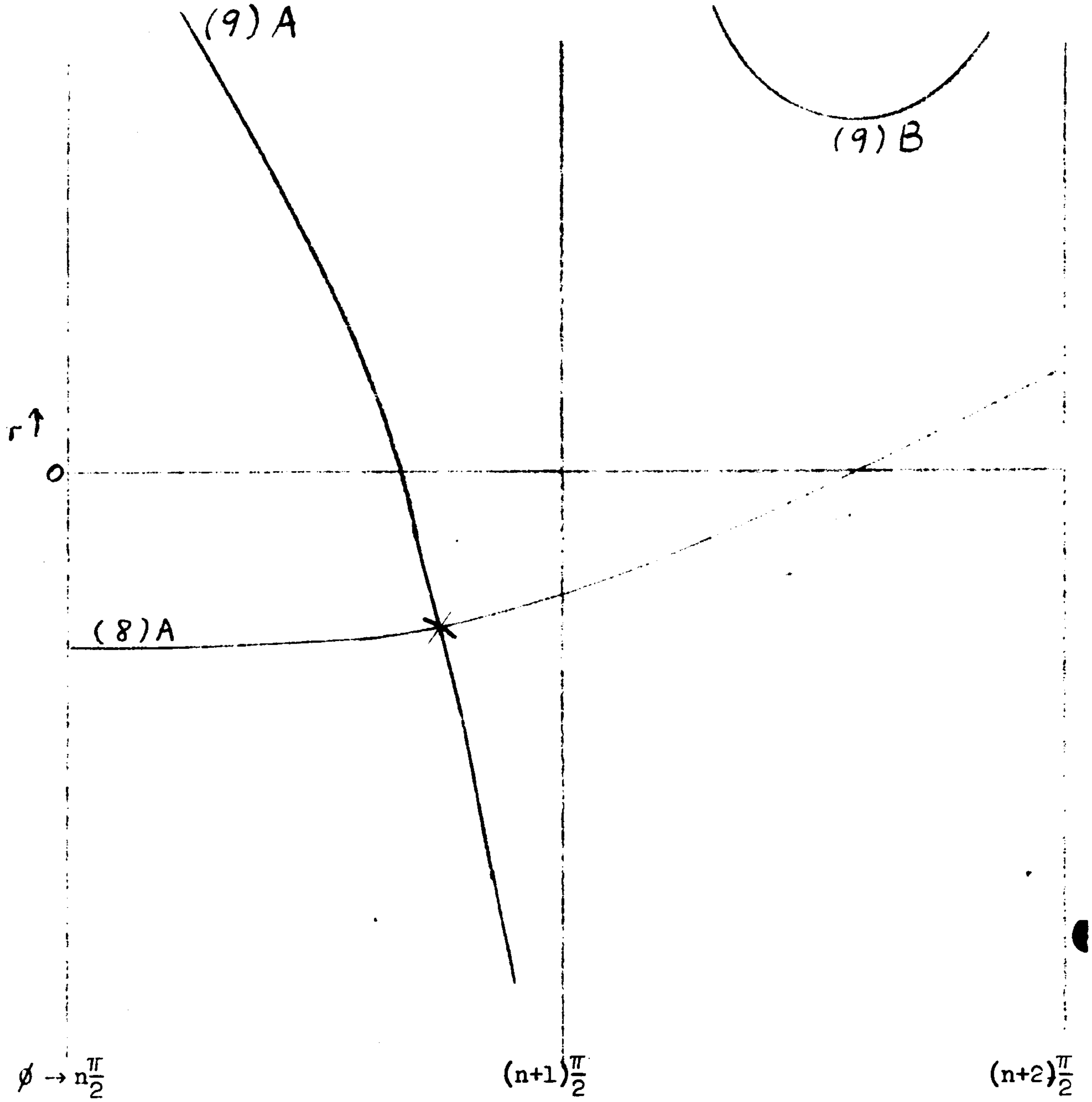


Fig. (12)

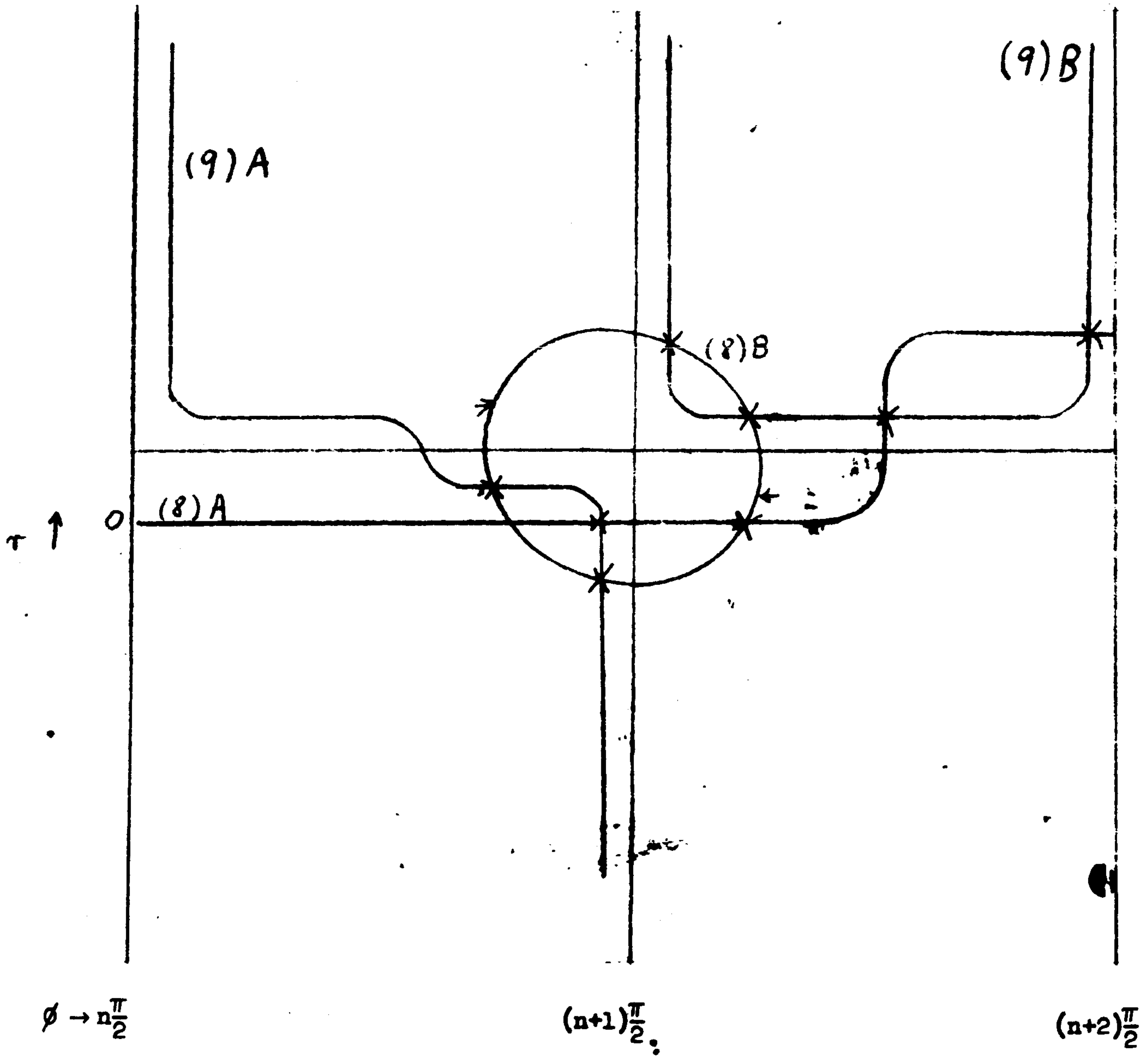


Fig. (13)

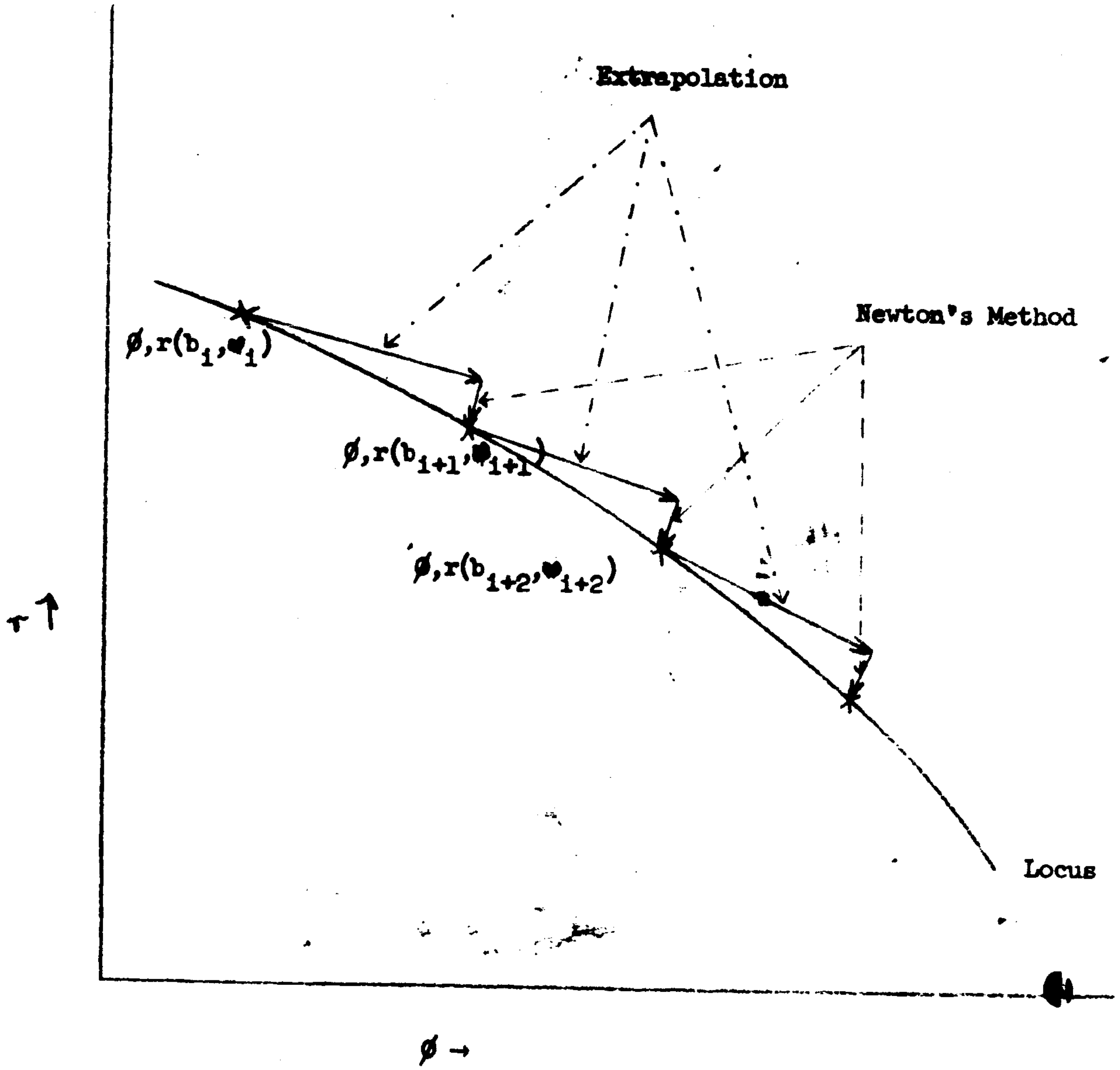

Extrapolation
Newton's Method
$\phi, r(b_i, \psi_i)$
$\phi, r(b_{i+1}, \psi_{i+1})$
$\phi, r(b_{i+2}, \psi_{i+2})$
Locus
r ↑
φ →

Fig. (14)

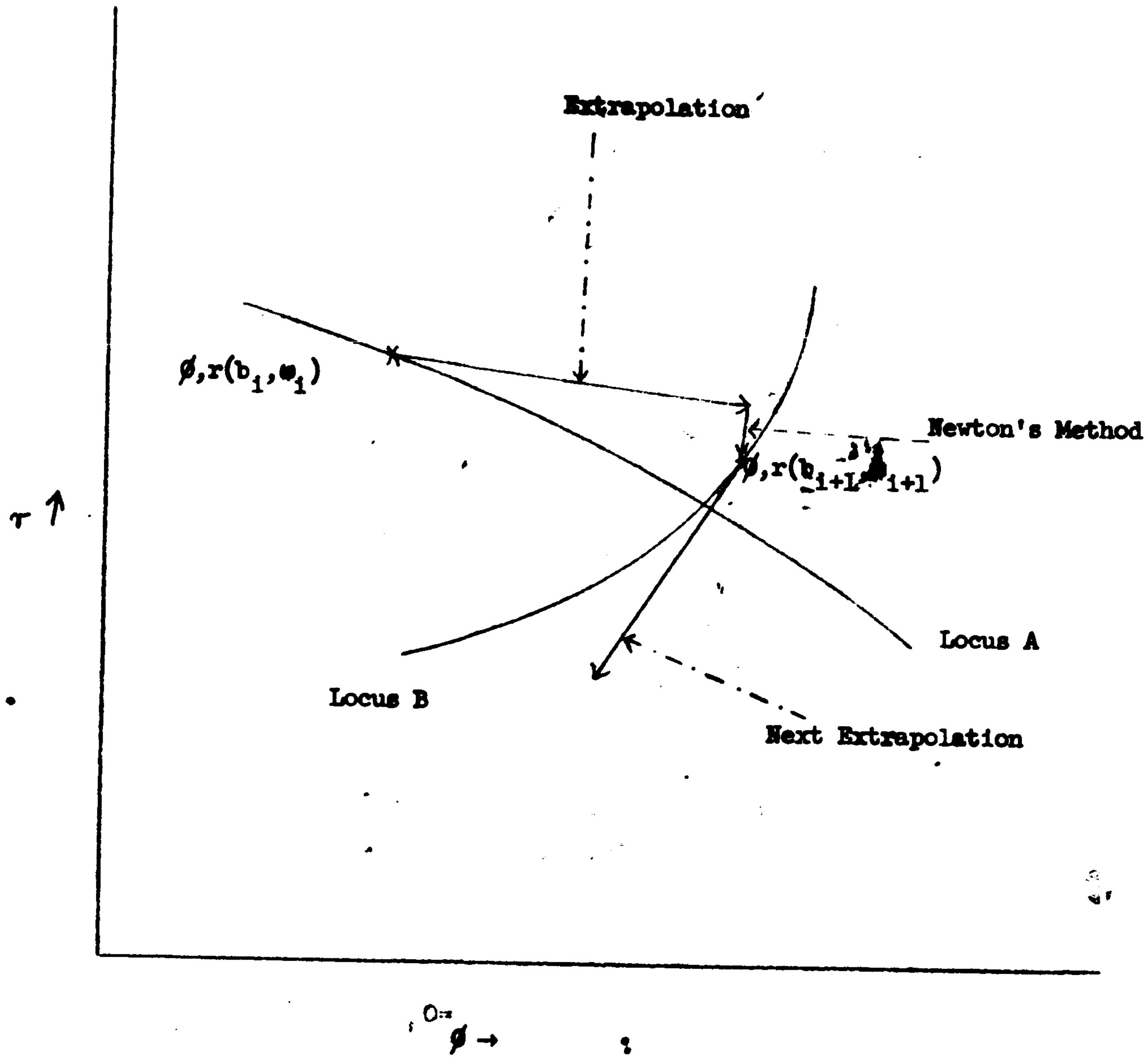

Extrapolation
ø,r(b_i,ω_i)
Newton's Method
ø,r(b_{i+1},ω_{i+1})
τ ↑
Locus A
Locus B
Next Extrapolation
ø →

Fig. (15)

inspection.

The technique then is to chose a canonical value of ω and a ratio of b to $\xi + f$ where ξ and f are kept constant and then using Newton's Method and the r, ϕ' arrived at by inspection determine exact solutions of (8) and (9). In N steps b and ω are incremented until they reach the values for which we actually wish to solve. With each of these steps the locus of the solution is followed by the procedure diagrammed in Fig. (14). At each point r, ϕ is extrapolated to give an approximate value and then Newton's Method is applied to correct this value.

What can go wrong is exemplified in Fig. (15) where the extrapolation can lead to a close approximation of the solution on another locus for that step. The ensuing step would locate a point on the wrong locus. To avoid this possibility a sufficient number of past extrapolations are retained so that it is always possible to extrapolate two steps ahead as in Fig. (16). Whenever it is noted that there is a discrepancy in the answer the extrapolated value is chosen. This accomplishes the double (1) test of determining that the extrapolations are valid prior to trouble so that the decision to have confidence in them can be made; and (2) later to select on the basis of this confidence the correct point. We will later see that this extrapolated value can be used in another way.

The procedure starts with a pair of solutions. The choice of such pairs is determined by the fact that if they disappear as solutions, the disappear simultaneously. In general, the two loci will converge together as in Fig. (17).

After converging, they will disappear. The test used here is that when the loci have reached a distance d less than three times the most recent extrapolation distance for the extrapolation of both, r, ϕ then the solutions are decided to be of no interest. (Of course, locus 8A is an exeption). How these solutions can disappear is sketched in Fig. (18) where it will be noted that eventually 8B and 9B no longer intersect. Also shown is how 8B will eventually disappear from the real plane altogether.

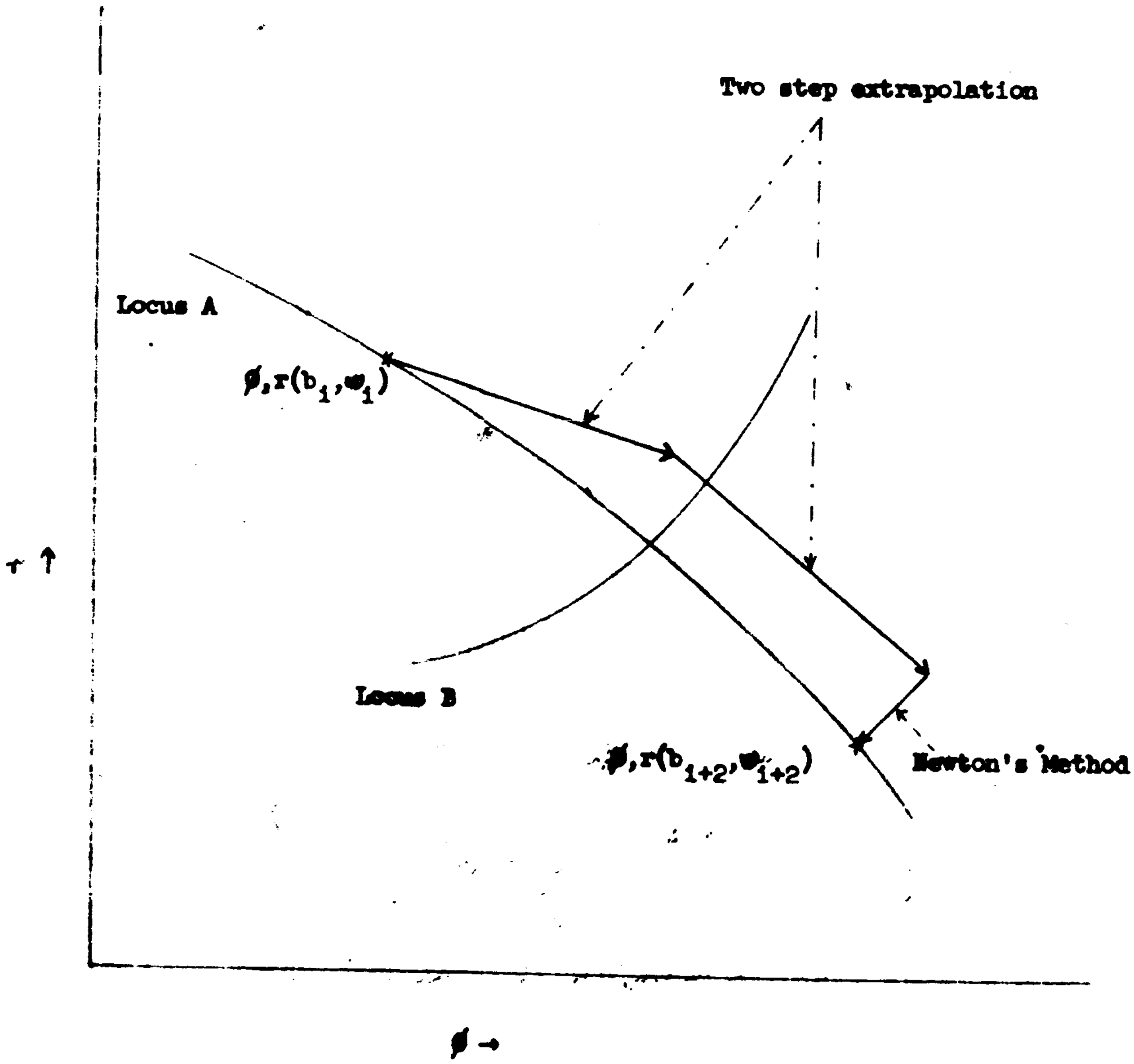

Two step extrapolation
Locus A
$\phi, r(b_i, \omega_i)$
r ↑
Locus B
$\phi, r(b_{i+2}, \omega_{i+2})$
Newton's Method
ϕ →

Fig. (16)

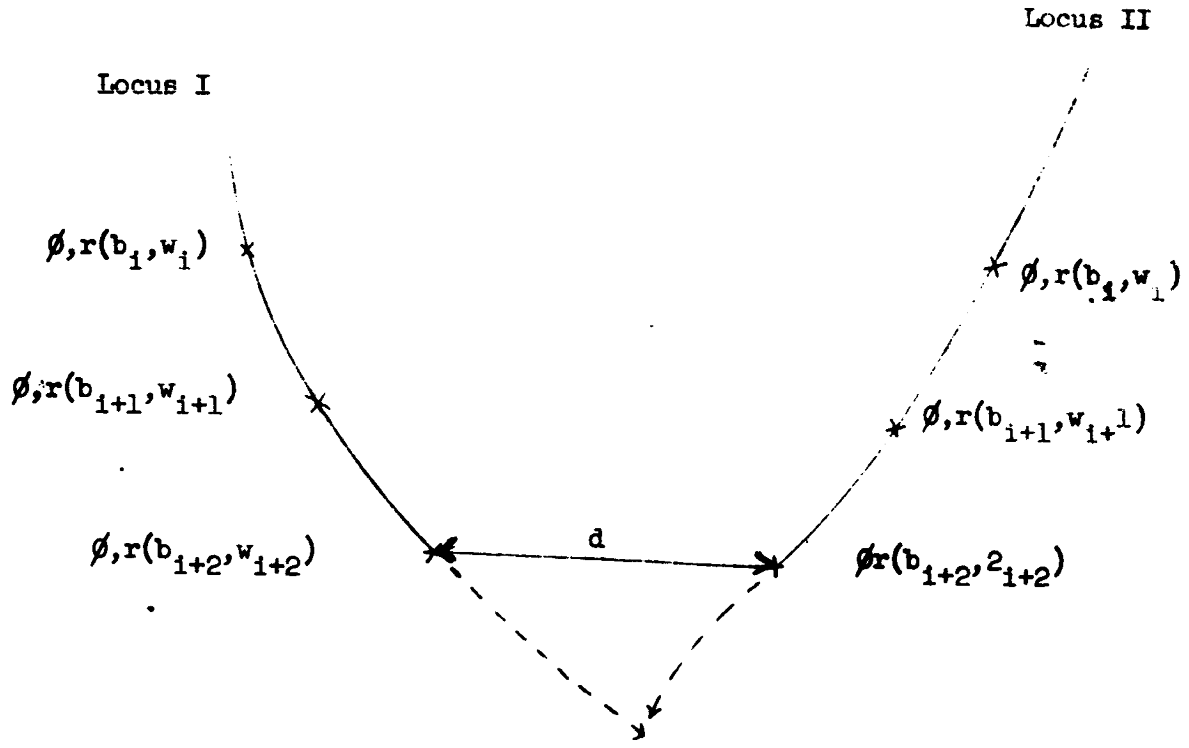

Locus I
Locus II
ø,r(b_i,w_i)
ø,r(b_{i+1},w_{i+1})
ø,r(b_{i+2},w_{i+2})
d
ø,r(b_i,w_i)
ø,r(b_{i+1},w_{i+}1)
ør(b_{i+2},2_{i+2})

Fig. (17)

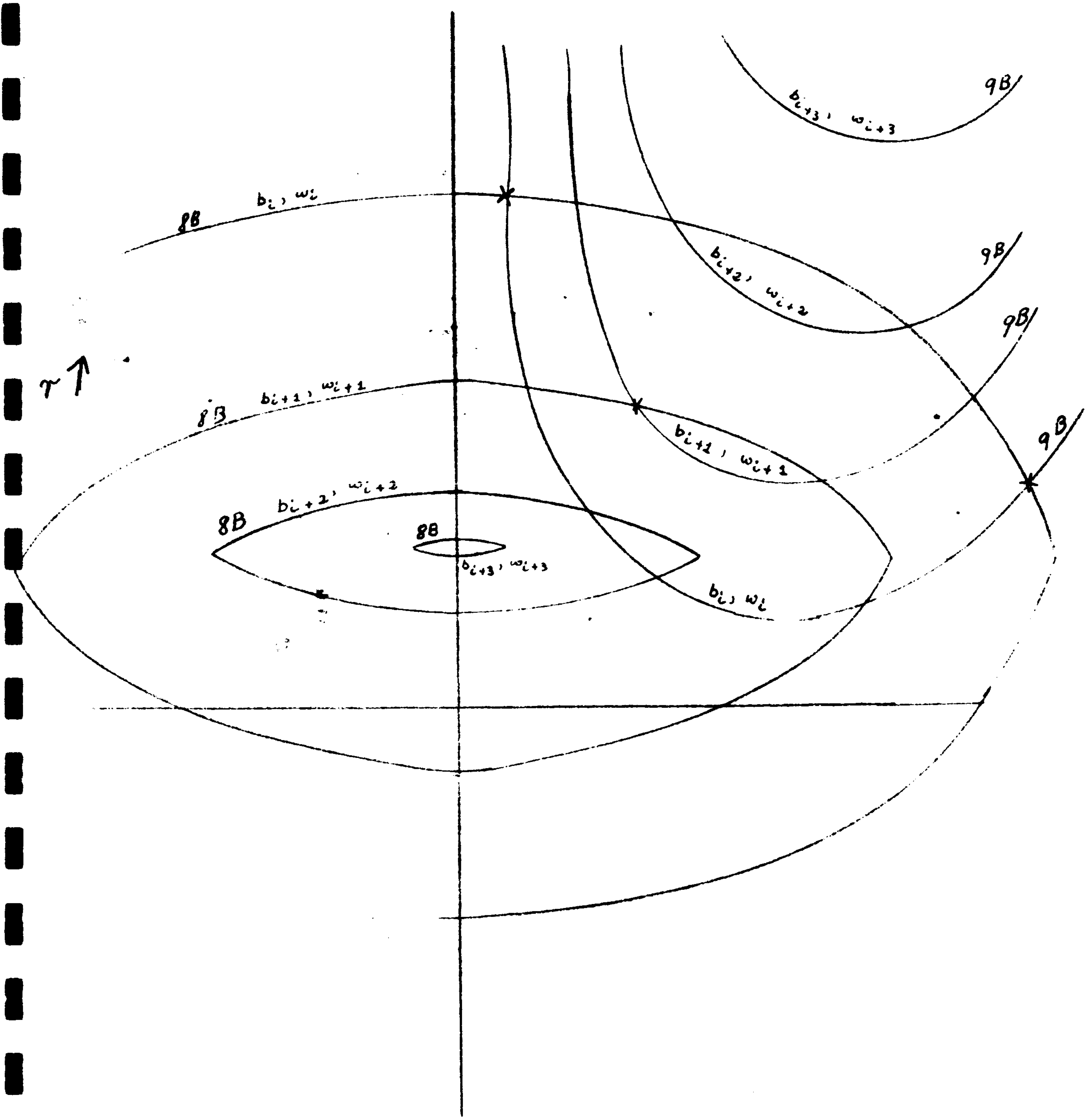

9B
b_{i+3}, w_{i+3}
8B
b_i, w_i
9B
b_{i+2}, w_{i+2}
9B
r ↑
8B
b_{i+1}, w_{i+1}
b_{i+1}, w_{i+1}
9B
b_{i+2}, w_{i+2}
8B
8B
b_{i+3}, w_{i+3}
b_i, w_i
ϕ →

Fig. (18)

Finally, it is possible that there will be no solution whatsoever in the case that ω takes on the value of an integral multiple of $\frac{\pi}{2}$. In this case Fig. (19) goes into Fig. (20) and the solution on the negative branch of 9A no longer appears.

This is essentially the procedure, the best distance is chosen at each step and the whole iteration resumes. In an Appendix the algebra of the algorithm which is actually pretty straightforward and an annotated Fortran Program for the polynomial version appear.

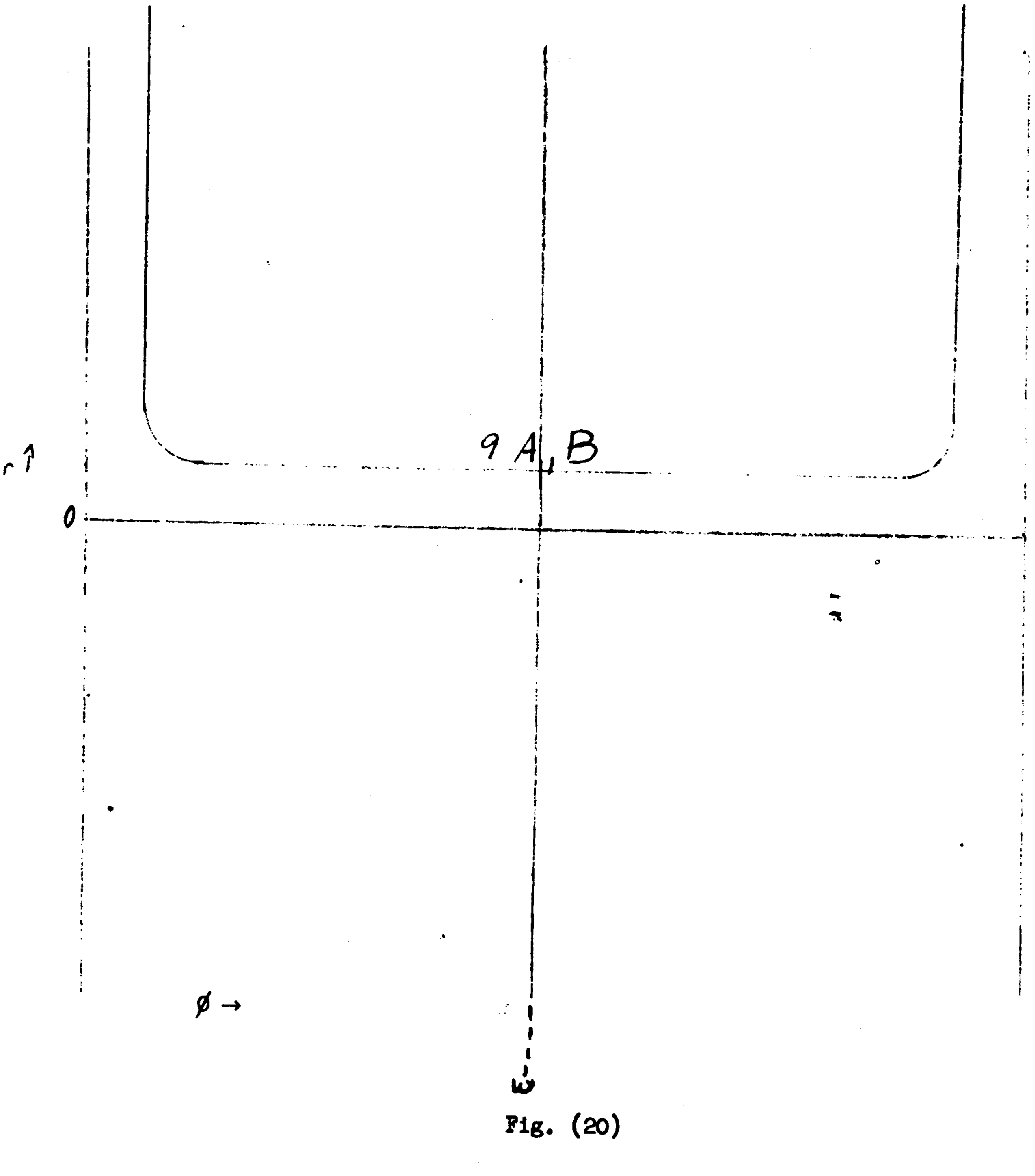


Fig. (20)

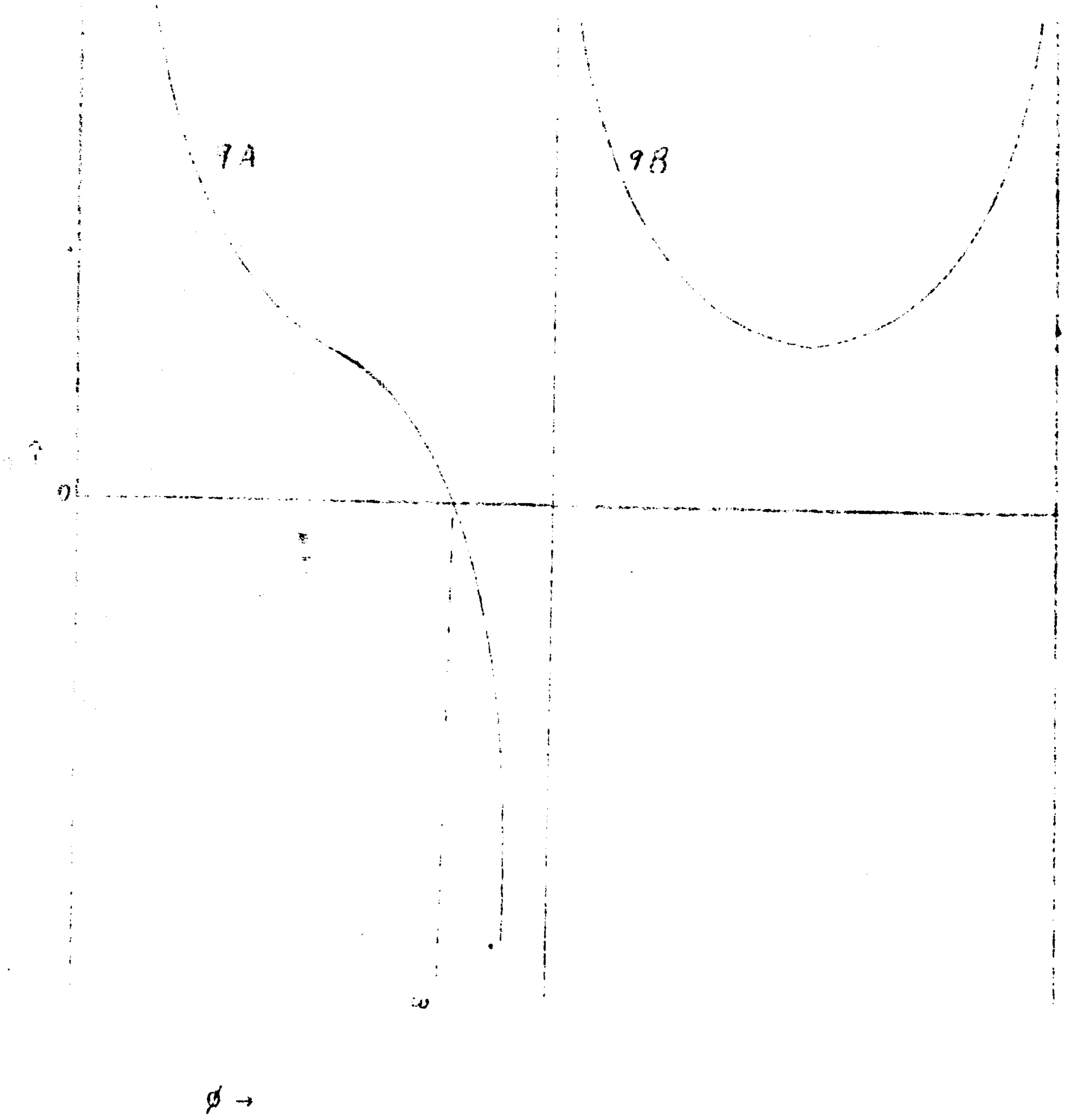


Fig. (19)

5.0 COMPUTROL SYSTEMS SYNTHESIZED

5.1 An Iterative Analog Computer that Solves the Euler-Lagrange Two-Point Boundary-Value Problem.

In the preceding section the theory of fuel optimal attitude control was applied to yield three algorithms which may be synthesized into computrol systems. This section gives the preliminary synthesis of these three control algorithms with due consideration also given to systems reliability, control actuator configurations, system transducers and remote command communication (system input/output).

5.1.1 Three On Line (On Board) Computers for Optimal Attitude Control.

In order to realize fuel optimal attitude control it was proven in the previous section that "bang-coast-bang" (+1, 0, -1) actuator signals are applied to the reaction jets for coarse control during large reorientations, then followed by linear control (actually regulation) whith set point biasing of the control moment gyros to hold at the origin. It also has been shown that there are three distinct methods of determining the switching signals that will optimally bring the vehicle to the origin. These are:

Method 1. Pre-computation by the adjoint-system method.

Method 2. Real time solution of the two-point boundary-value problem.

Method 3. Approximate closed form solution of the Hamilton-Jacobi partial differential equations.

This section gives the initial logical design of three digital computers which respectively realize the three aforementioned methods of determining the optimal switching trajectories. These are:

1. The Stored Function Computer, which constitutes a table look up of the switching signals as a function of where the vehicle is in state space, the function (algorithm) determined by Method 1.

2. The Relaxation Computer, which gives a powerful yet basically redundantly simple computer realization of the two-point boundary-value problem (Method 2) solved faster than real time.

3. The Closed Form Computer, which is inherently the simplest realization of optimal control since the computer is derived from Method 3, the approximate closed form of the Hamilton-Jacobi equations.

Potentially all three of these computers could be employed in a final attitude control system because of the powerful interplay that exists among them for producing adaptivity and reliability (redundency of different kind). However, each computer (or method) is also potentially complete unto itself. An extremely important result of this study will be the answer to the trade-off possible in the final control computer design.

5.1.1.1 The Stored Function Computer. Fig. I shows a block diagram of the stored function computer. The steps required to arrive at the final form of this computer include the computation (throughout phase space) of the value that the switching function should have an adequately dense set of points. From this information items are determined. The first of these items is the form of the information that must be designed into the computer as permanent memory in what would normally be regarded as an addressable computer store. The other important item required is information for the design of a decoding net that by directly setting the actuator values for a large part of phase space will reduce materially the size of the store.

To make clear what kind of a computer is involved turn to Fig. I. Coordinate pulses arriving from the analog to digital converters are fed into the Counters. Three Counters are shown here. For the attitude control computer six Counters will be needed to describe the six important state variables. Scaling of the rate at which the pulses arrive can be done by adjustment of the analog to digital converters.

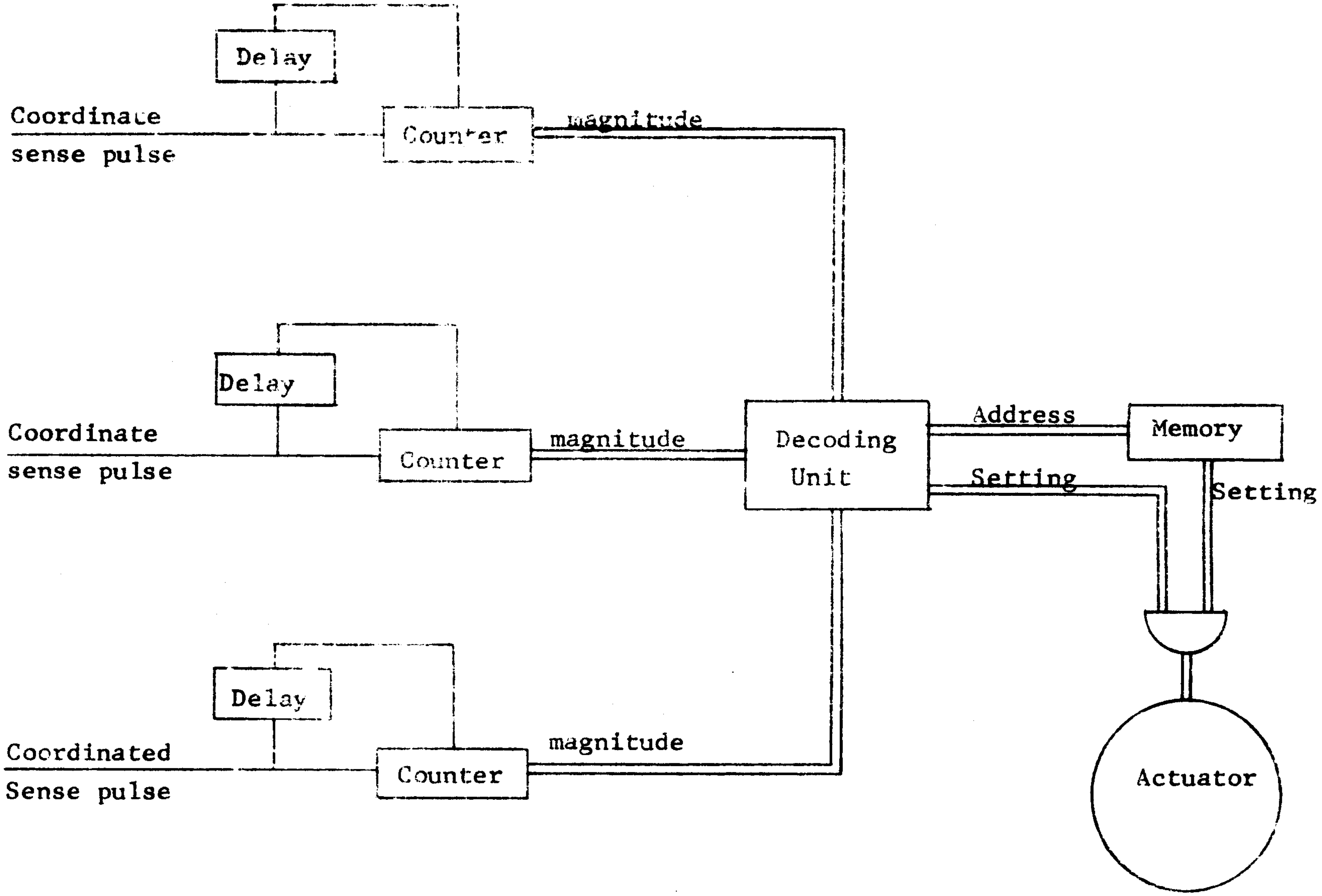


Fig. I - Stored Function Computer

This scaling is the feature that makes the computer input simple. The operation is of the following form. Pulses are constantly sent to the Counter at a rate proportional to the reading of the Sensor. As they arrive they are added to the Counter. The same pulses are entered onto a Delay Line. At the point of exit from the Delay Line they are subtracted from the Counter. Thus the Counters always contain the number of pulses in the Delay Line.

Correction of the Counters to avoid accumulation of errors can be carried out by periodically clearing the Delay Line and the Counter simultaneously. Recalibration of the whole input system can be accomplished by the sensing of phase space points corresponding to Counter carry points and perturbing the Sensor adjustments in a manner depending upon whether the carry is sensed before or after the check point is sensed.

The output of the Counters, which constitute a point in quantitized phase space, is fed into a Deconding Net. The Decoding Net will either determine a Memory address or an Actuator setting. The Memory address is determined in the case that the elaboration of the Decoding Net is not sufficient to provide the correct Actuator setting.

As the Actuators can be set into any of 27 states five bits of information are required. It is estimated that a maximum bound of storage for the correct settings of a million point in phase space will be required (5 million bits) and this information will densely fill phase space when full advantage is taken of the symmetries (rotations and reflections) that exist. Present day reliable large volume (information) stores of this order are in existence that require less than 1 cubic foot and weigh less than 75 pounds. The useful form of these five bits of information per point in phase space is in the form of three sets of triple valued states. That is, each of the three Actuators should be either full on in either of two directions or else full off.

The simplest version of this machine would be to have the settings for all of the points in phase space stored in the Memory and to have the outputs of the Counters directly determine the address to be referenced in the Memory. Though the simplicity of this computer is certainly appealing, the size of the Memory may still be reduced to make the computer minimal. Large areas of phase space have the same setting and this information can be directly incorporated into the switching circuitry of the Decoding Net. Full examination of the switching function throughout the phase space will give a clear understanding of the best possible match of switching circuitry and Memory so as to reduce the volume and weight of the total machine.

5.1.1.2 The Relaxation Computer. The two-point boundary-value problem of the direct and adjoint systems is best solved faster than real time utilizing the relaxation algorithm given above. Furthermore by using an Aeronca proprietary computer* organization the computation can take place faster than real time with basically slow-speed, simple computing units. The reason this is possible is that a multiplicity of the simple computing units are simultaneously working in parallel under one program control. Inherent in this parallel organization of simple units is not only the intrinsic reliability because of the units themselves (because of slow speed) but also the increased system reliability through the use of redundancy at a functional level. With this approach, when a unit is detected to be malfunctioning it is switched out and the load is handled by the remaining units. This can continue until there are not enough units left to handle the relaxation computation in the time required. Further, when this has occurred, the remaining unit can be caused to behave like the closed form computer discussed in the next section.

Before the parallel organization of the simple units for solving the relaxation algorithm is discussed, the simple unit will be explained. The unit is of the form of the simple computer shown in Figure I. Alpha and Beta are two seperate banks of

*See Aeronca Report 12-1 entitled "Parallel Systems Organized Computer" (PSOC).

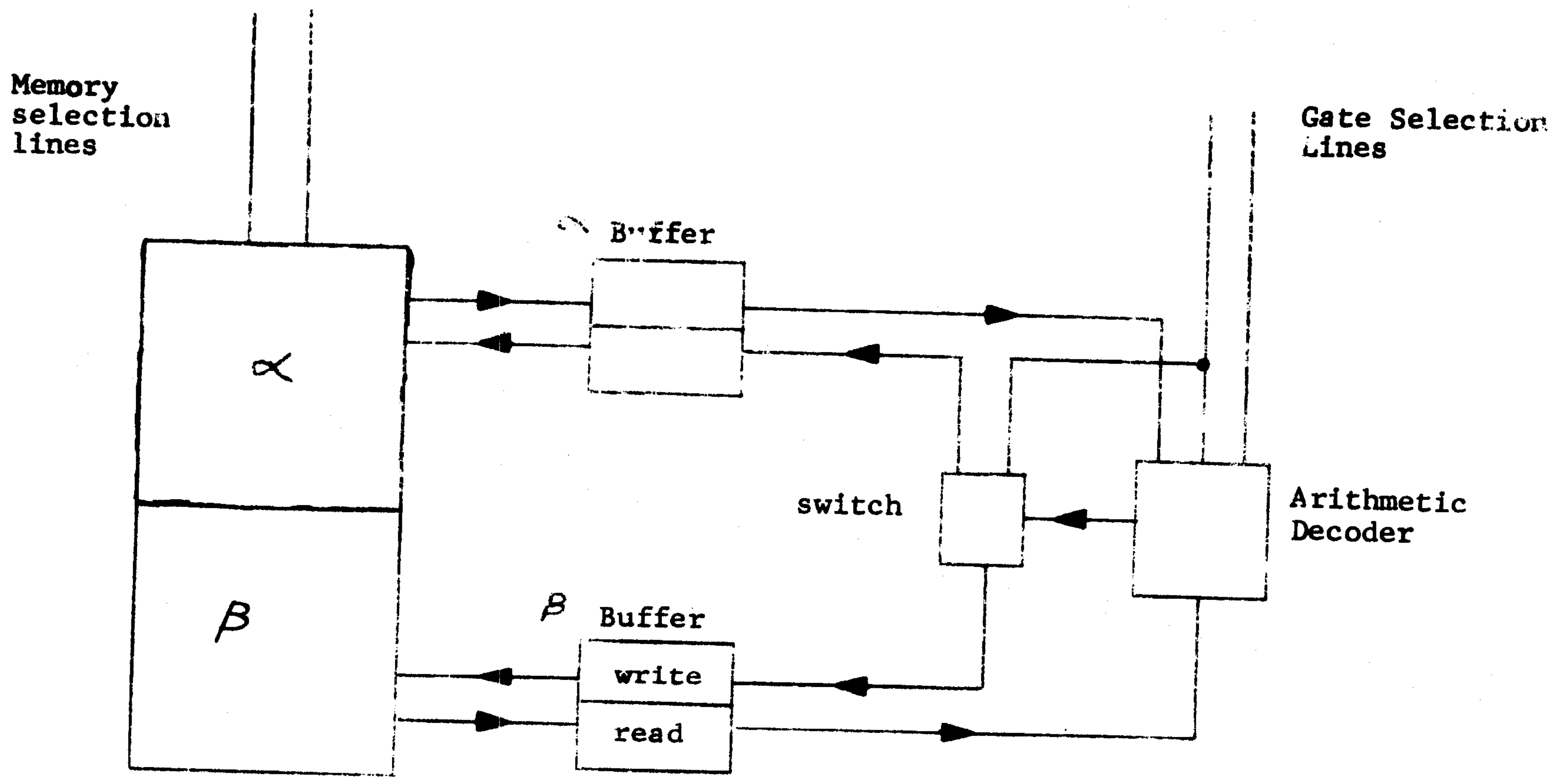

Memory
selection
lines
Gate Selection
Lines
Buffer
switch
Arithmetic
Decoder
α
β
β Buffer
write
read

Fig. I - Unit

memory that can be read or written into simultaneously from the buffers. The buffers are the inputs to an arithmetic decoder, whose function is determined by the remaining inputs from the gate control lines. Two other inputs to the Arithmetic Decoder are from the internal state memories. The first of these is a carry state and the second is another "on-off" state. One of the gates is set to give the output of the Decoder to a hub of either the alpha or the beta Output Buffers. At strobe limit this value is used to set the Output Buffer. The other Output Buffer is set to write back the information that was in the corresponding Read Buffer. From the simplest viewpoint the Unit should be regarded as a simple serial two address computer. The purpose of the two addresses is to determine the two operands, one being obtained from each of the two memory banks. The result is then returned to one of the memory banks. However, there is no program stored in this computer as there is no provision for the decoding of operations. The choice of the operands and the memory locations is determined by the values of the Gate Lines and Address Lines coming from external decoders. A rudimentary programming facility does exist in the machine in the form of the possibility of the machine being able to take one of two states. The setting of this state in turn modifies the setting of the gates.

This absence of the complete facility of the computer is by design, the Unit being one of many Units organized together as in Figure II; where to the right of the Diagram are dotted line block units. It is the configuration of this diagram that is proposed for the relaxation computer. To the left is a Stored Program Memory that has the steps of the program that must be carried out by the computer in the solution of the algorithm. The successive steps of the Stored Program are selected by the Stored Program Counter. Each of the Stored Program instruction is decoded by the Gate Selection Switch, the Information Shift Control, and the Memory Selection Switch.

First note that the Gate Selection Switch and the Memory Selection Matrix open the same gates and select the same addresses in every Unit. The Information Shift

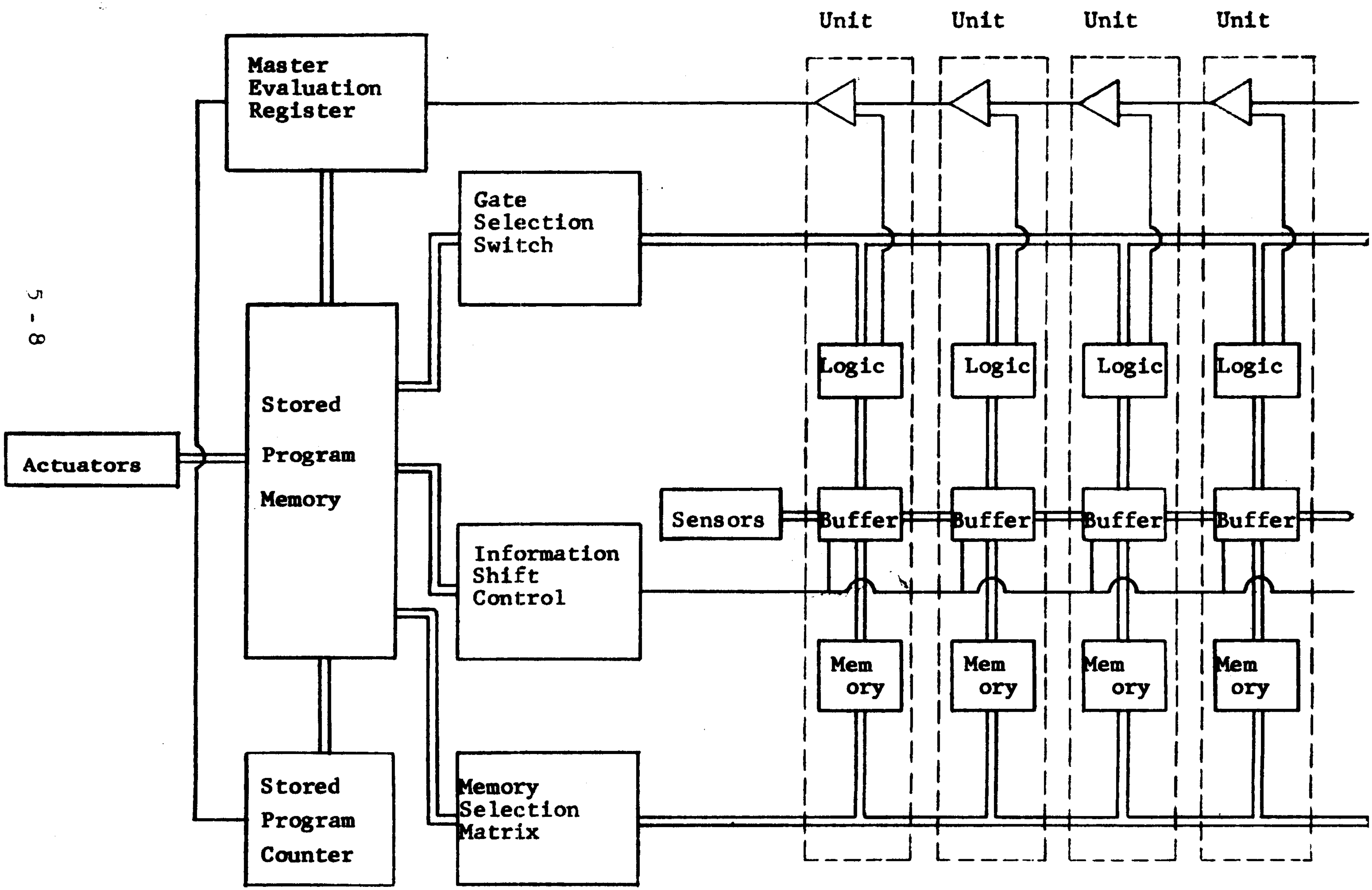


Fig. II - Relaxation Computer

Control also gives the same information to each of the units. In effect the operation of each of the units is exactly the same and the sole effect determining a different operation of the different units is that they each have different information stored in them. Thus it is immaterial how many Processing Units are in the machine, the operation being indifferent to the length of the string of processors to the right.

One point not brought in Figure I but designated in Figure II is that the Buffers are connected together as shift registers, that is the Alpha Buffers form a shift register the length of the Unit chain and the Beta Buffers form a shift register again the length of the chain of Units. Information is uniformly transferred from Unit to Unit by means of these shift registers which operate independently. The control of these shift registers is what is determined by the Information Shift Control. Note that the sensors are connected to this information channel.

A sequence of "OR" gates at the top of the diagram provide the channel by which information is collected from the processing units to be transferred to the Master Evaluation Register. Information in the Master Evaluation Register is consulted to determine the decisions that must be made by the stored program. The consultation is generally in the form of inquiring as to whether a stored constant is exceeded by the contents of this register and an affirmative answer is manifested in the form of a decision to instruct the stored program counter to accept the next stored program constant as a new setting.

The remaining anomaly, namely that the information from the Units is "orred" together, and thus would seem to be non-unique in its form when being transferred to the Master Evaluation Register, is resolved on realizing that the logic of the Unit contains a state setting. At the time of the transfer of the information the setting of the states of all of the units save one is such that information will not be transferred out of them. Thus a unique signal of a single Unit is read by the Master Evaluation Register. Which Unit is on at the time of transfer of information is determined by internal computations carried out by the seperate units.

Computation of the relaxation algorithm proceeds in the following manner. Each of the Units is identified with a point in the quantization of the trajectory. The Computation of the corrected value of the coordinate and adjoint coordinate values is computed simultaneously in each unit. As the previous values of the nearby points are needed they are obtained by uniform transfer of information through the shift registers. At the end of each iteration a computation is done to determine the value of the differences of the last two iteration and all of the Units save one carrying the largest value of this number are turned to the "off" state. This sole Unit that is left "on" transfers its information to the Master Evaluation Register so that during the course of the next iteration it can be decided whether or not the computation has bee completed. When the computation has been completed and the information needed to determine the setting of the Actuators is transferred to the Master Evaluation Register where it is used to determine the setting of the Actuators. The selection of the number of Units that must be "flown" can always be optimized on the basis of necessities of the mission. In particular, a given number of real units can be multiplexed to provide an interger multiple of this number of virtual units. With a multiple identification of points with units the important feature of rapid propagation of the boundary-value effects can be retained if an identification scheme such as that outlined in Figure III is used.

5.1.1.3 Closed Form Computer. The closed form computer has the characteristics of the general digital computer with a restriction of its characteristics to only those needed for the solution of the problem. In particular, in that the computation is based only upon the present value of a set of sensors, no large store of information is required. The data that is kept has only the purpose of providing heuristic checks. Stored logic will suffice for the program since only a single program need be considered. The magnitude of the accuracy will be a function of the fact

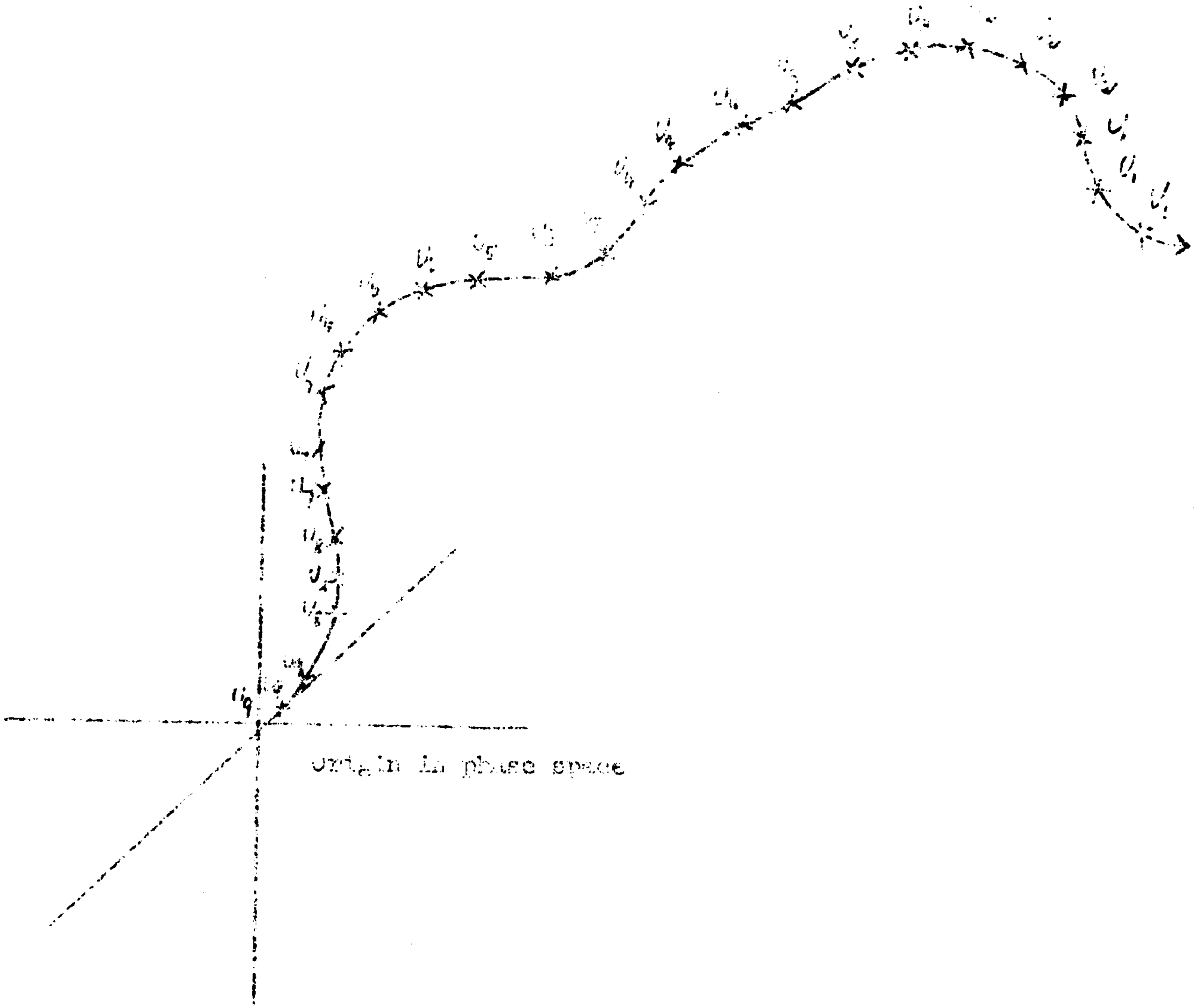


Figure III.

Unit Identification with Points of Discretized Trajectory.

that the control is only an approximation to optimal control and the computer need be no more accurate than the approximation. The control must be designed to carry out certain operations with optimal efficiency. In this particular computer the arcsine and the square root must be efficiently computed. This entails all of the registers for rapid division process as the arcsine is best obtained by a continued fraction expansion and the square root by a more general use of the division hardware. The basis for this computer will be the simple Unit discussed in the previous section with the addition of the special commands above.

6.0 EXAMPLES OF COMPUTROL SYSTEMS.

6.1 The Control of a Rotating Servomechanism.

In the effort to derive a switching characteristic applicable to practical transducers in a medium power servomechanism, the following calculations have been made.

Consider the following configuration for motor and load, in which the following idealizations have been assumed:

a. Motor transfer function is linear and second order.

b. Gear train is free of backlash.

c. Load parameters are constant.

d. External torque variations never exceed the stall torque of the motor.

See Figure 1.

If the servo torque-speed curves have the following forms

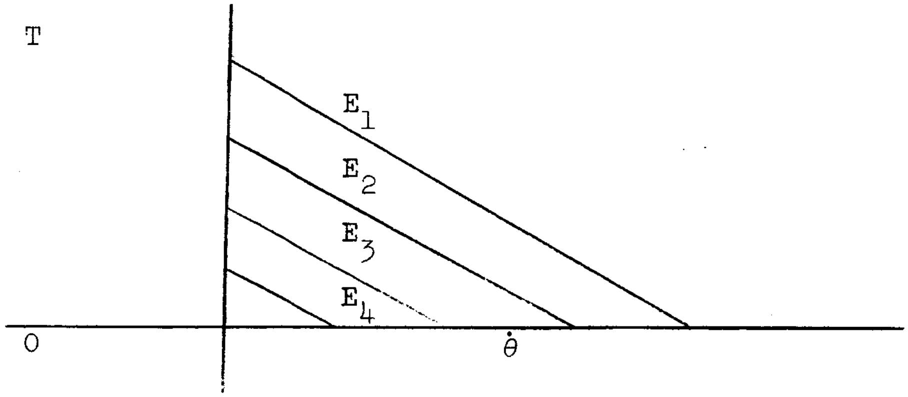


Figure 2

end $E_1 > E_2 > E_3 \ldots$

and since

$$T = \frac{\partial T}{\partial E} A + \frac{\partial T}{\partial \theta'} \dot{\theta}'$$

then let

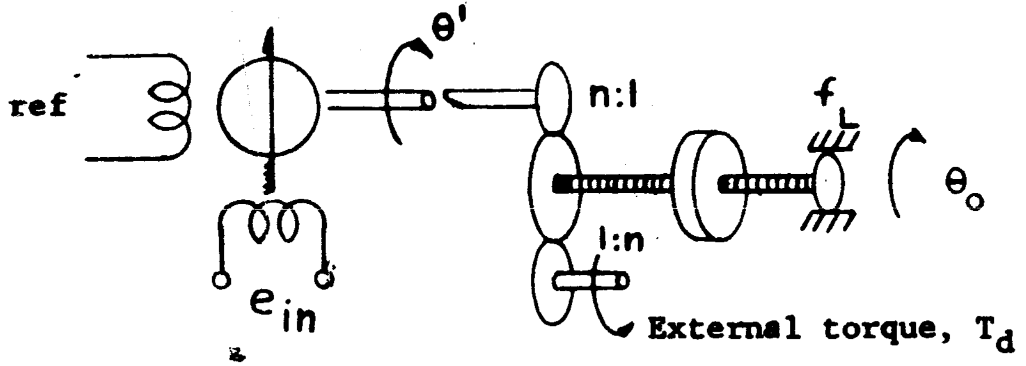


Figure 1

Pictorial of Motor and Load Configuration

$$\frac{\partial T}{\partial E} = K_e ; \qquad \frac{\partial T}{\partial \theta'} = -K_n \quad .$$

Hence the developed torque

(1) $$T_m = K_e\, E_{in} - K_n \dot{\theta}' \quad .$$

The load torque at the motor shaft

(2) $$T_L = \left(\frac{J_L}{n^2} + J_m \right) \ddot{\theta}' + \left(\frac{f_L}{n^2} + f_m \right) \dot{\theta}' - \tau_d \quad .$$

Equating (1) and (2) we have the system of differential equations

(3) $$\left(\frac{J_L}{n^2} + J_m \right) \ddot{\theta}' + \left(\frac{f_L}{n^2} + f_m + K_n \right) \dot{\theta}' = K_e\, E_{in} + \tau_d .$$

Let

$$\left(\frac{J_L}{n^2} + J_m \right) = J, \qquad \frac{\tau_d}{n} = T_d$$

$$\left(\frac{f_L}{n^2} + f_m + K_n \right) = f$$

$$\theta' = n\theta$$

$$\frac{K_e E_{in}}{n} = F \operatorname{sgn} \sigma \qquad \{\text{the extremal constraint}\}$$

Hence (3) becomes

(4) $$J\ddot{\theta} + f\dot{\theta} = F \operatorname{sgn} \sigma + T_d$$

If the solution of this equation is to follow a given input, $\theta_i(t)$ so chosen that

(4a) $$\theta_i(t) = a + bt, \qquad |b| < \frac{F+T_d}{f} \quad \text{(runaway velocity)}$$

within any finite interval, then it is convenient to define a new variable, $e(t)$, the actuating error

(5) $$e(t) = \theta_i(t) - \theta(t))$$

clearly $|e(t)| \to 0$ as $\theta(t) \to \theta_i(t)$.

After substitution of (5) and (4a){ (4) becomes

(6) $$J\ddot{e} + f\dot{e} = (fb - T_d) - F \operatorname{sgn} \sigma \ .$$

We then desire that the choice of σ which will bring $e(t)$ and its derivative to the origin from any initial position $(e_o, \dot{e}_o)$.

Let us solve (6).

Set

$$e(0) = e_o$$
$$\dot{e}(0) = \dot{e}_o \quad .$$

Now

$$e(t) = A + B\epsilon^{-\frac{f}{J}t} + \frac{1}{f}[(fb - T_d) - F \operatorname{sgn} \sigma]t$$

$$\dot{e}(t) = -\frac{f}{J}B\epsilon^{-\frac{f}{J}t} + \frac{1}{f}[(fb - T_d) - F \operatorname{sgn} \sigma]$$

then

$$e_o = A + B$$
$$\dot{e}_o = -\frac{f}{J}B + \frac{1}{f}[(fb - T_d) - F \operatorname{sgn} \sigma]$$

hence

$$B = \frac{J}{f^2}\,[(fb - T_d) - F \text{ sgn } \sigma] - \frac{J}{f}\,\dot{e}_o$$

$$A = e_o + \frac{J}{f}\,\dot{e}_o - \frac{J}{f^2}\,[(fb - T_d) - F \text{ sgn } \sigma] \;.$$

Finally

$$(7) \quad e(t) = e_o + \frac{J}{f}\,\dot{e}_o - \frac{J}{f^2}\,[(fb - T_d)\; F \text{ sgn } \sigma] + \{\, \frac{J}{f^2}\,[(fb - T_d) - F \text{ sgn}\sigma] - \frac{J}{f}\,\dot{e}_o\}\epsilon^{-\frac{f}{J}t}$$

$$+ \frac{1}{f}\,[(fb - T_d) - F \text{ sgn}\sigma]t$$

and

$$(8) \quad \dot{e}(t) = \frac{1}{f}\,[(fb - T_d) - F \text{ sgn}\sigma] - \frac{1}{f}\,[(fb - T_d) - F \text{ sgn}\sigma - f\dot{e}_o]\epsilon^{-\frac{f}{J}t} \quad .$$

If we eliminate the parameter, t, the resultant equation defines two families of phase plane trajectories, where the phase plane is defined to be $e(t)$, $\dot{e}(t)$ plane. These two families differ in the assumed algebraic sign σ. See the following sketches wherein it is assumed that b, the input rate, and T_d, the disturbing torque, are held constant throughout the trajectory.

Note that only that part of the trajectory which is heavily shaded in the sketch, will bring the system to the origin with no torque reversal. This then is the desired final trajectory. Let us solve equations (7) and (8) for this trajectory by eliminating t, and setting the endpoint equal to the origin, $e(t) = \dot{e}(t) = 0$. Hence, (8) becomes

$$(9) \qquad \epsilon^{-\frac{f}{J}t} = \frac{\frac{1}{f}\,[(fb - T_d) - F \text{ sgn}\sigma]}{\frac{1}{f}\,[(fb - T_d) - F \text{ sgn}\sigma] - \dot{e}_o}$$

and

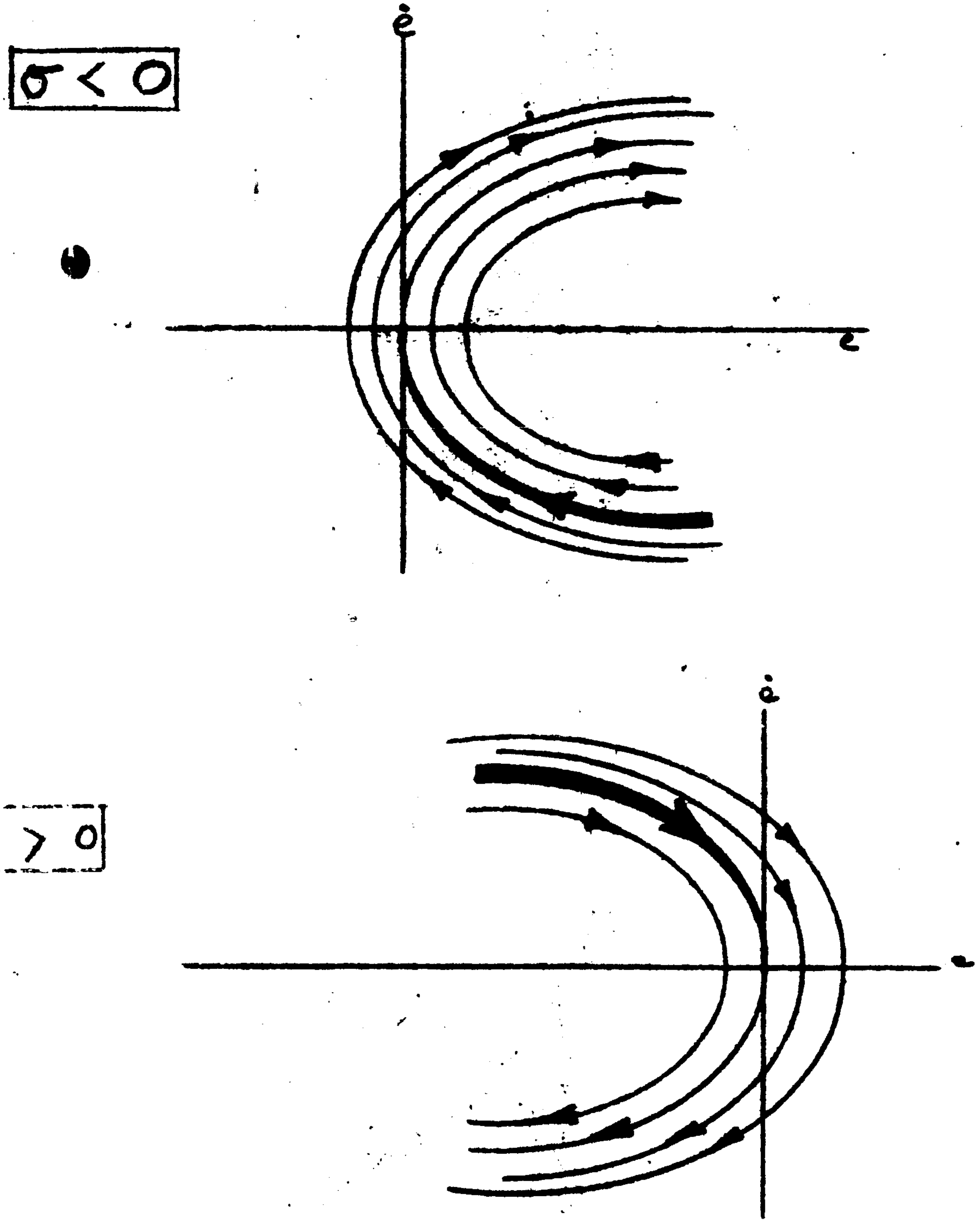


Figure 3.

Phase-plane semi-trajectories

(9a) $$t = \frac{J}{f} \log \left\{ \frac{\frac{1}{f}\,[(fb - T_d) - F\,\mathrm{sgn}\sigma] - \dot{e}_o}{\frac{1}{f}\,[(fb - T_d) - F\,\mathrm{sgn}\sigma]} \right\} .$$

Clearing (7) and (8) we have

(10) $$0 = e_o + \frac{J}{f}\,\dot{e}_o + \frac{J}{f^2}\,[(fb-T_d) - F\mathrm{sgn}\sigma] \log \left\{ \frac{[(fb-T_d) - F\mathrm{sgn}\sigma] - f\dot{e}_o]}{[(fb-T_d) - F\,\mathrm{sgn}\sigma]} \right\}$$

Equation (10) is a double valued implicit function of $(e_o, \dot{e}_o, \sigma)$.

See illustration on following page.

Note that in the I and III quadrant solutions, it is implied by equation (9a) that we reach the origin in negative time, i.e., an unrealizable solution.

In other words, we require that equation (10) holds in reality only if

(11) $$\mathrm{sgn}\sigma = \mathrm{sgn}\ e$$

Hence

(12) $$0 = e_o + \frac{J}{f}\,\dot{e}_o - \frac{J}{f^2}\,[(T_d-fb) + F\mathrm{sgn}\ \dot{e}_o] \log \left\{ 1 + \frac{f\dot{e}_o}{[(T_d-fb) + F\mathrm{sgn}\ \dot{e}_o]} \right\}$$

becomes the optimal trajectory, and indeed it will be shown, is the switching curve.

If in the derivation of (12) we let the endpoint be given

$$e(t) = e_1$$

$$\dot{e}(t) = 0$$

then

(12a) $$e_1 = e_o + \frac{J}{f}\,\dot{e}_o - \frac{J}{f^2}\,[(T_d-fb) + F\mathrm{sgn}\ \dot{e}_o] \log \left\{ 1 + \frac{f\dot{e}_o}{[(T_d-fb) + f\mathrm{sgn}\ \dot{e}_o]} \right\} .$$

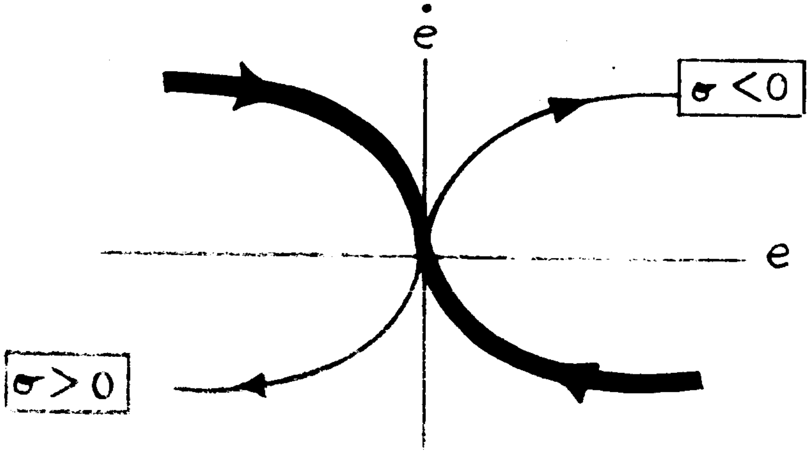


Figure 4.

Second Order Switching Boundary

which implies that the system comes to rest with an overshoot of e_1. Note that the sign of e_1 is also the sign of the applied torque which will cause the system to evolve to a point of interception with the ideal trajectory (12).

Hence, if we set $e_1 = \sigma$ and the applied torque equal to, $F \operatorname{sgn} \sigma$, then we generate the complete closed loop ideal switching characteristic for arbitrary initial conditions.

$$\sigma = e_o + \frac{J}{f}\dot{e}_o - \frac{J}{f^2}[(T_d - fb) + F\operatorname{sgn}\dot{e}_o] \log \left\{1 + \frac{f\dot{e}_o}{[(T_d - fb) + F\operatorname{sgn}\dot{e}_o]}\right\} \tag{13}$$

For step response only, with no torque adaptation we have

$$\sigma = e_o + \frac{J}{f}\dot{e}_o - \frac{JF}{f^2}(\operatorname{sgn}\dot{e}_o) \log\left(1 + \frac{f\dot{e}_o}{F}\right) \ . \tag{14}$$

See illustration on following page.

Equations (13) and (14) are completely rigorous with respect to the assumptions listed at the beginning of this section; however, we note the sensors of a rotating servomechanism are really measuring the variables of a congruence class, namely

$$\theta_i = \theta + 2\pi n \qquad (n = \text{any integer})$$

is a solution of the equations.

Representing $\theta \equiv \theta \bmod 2\pi$ is equivalent to mapping the phase place on a cylinder. See Figures 5, 6, 7 and 8.

The construction of a computer analog of this switching criterion is non-trivial. Several structures present themselves, however. Perhaps the most generally useful method is the analog synthesis.

This difficulty with the multipli-connected phase surface is avoided in part by the selection of components.

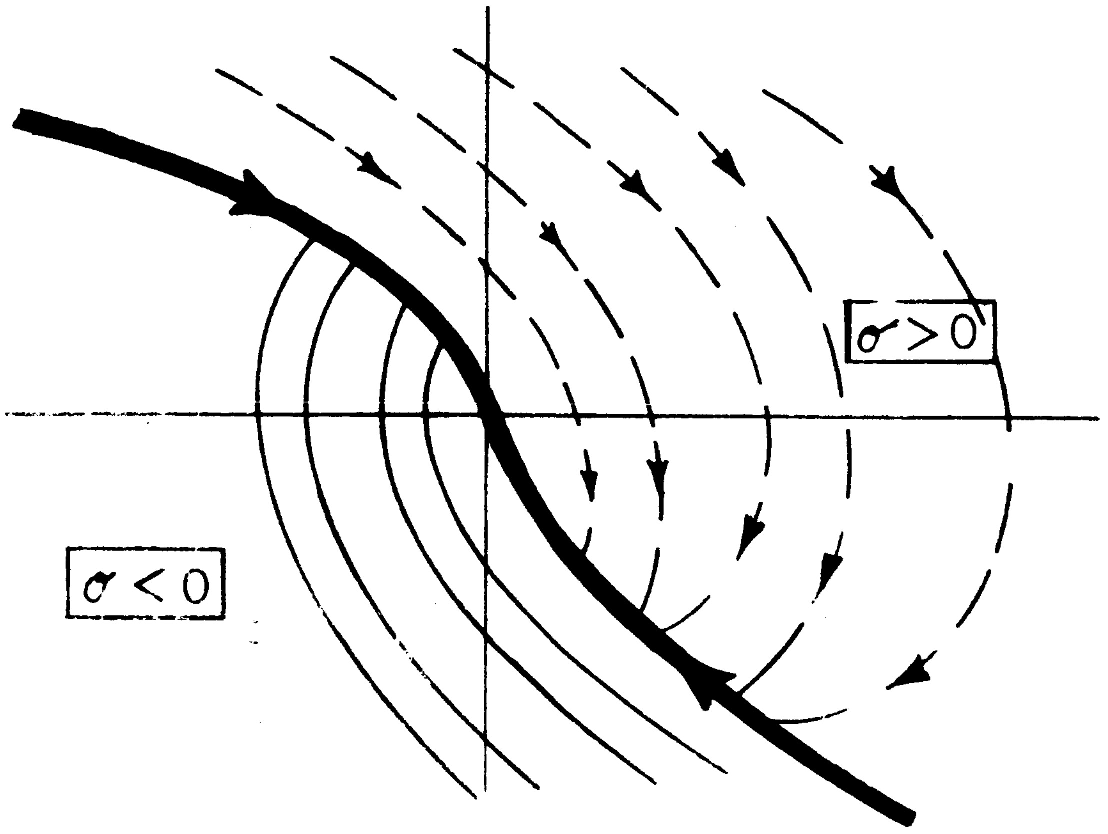


Figure 5.

Phase-plane Trajectories

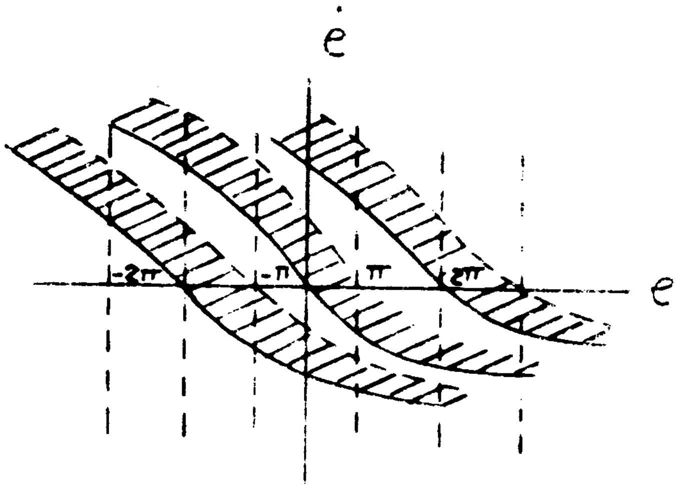


Figure 6.

Cartesian Map of Ambiguous Phase-plane.

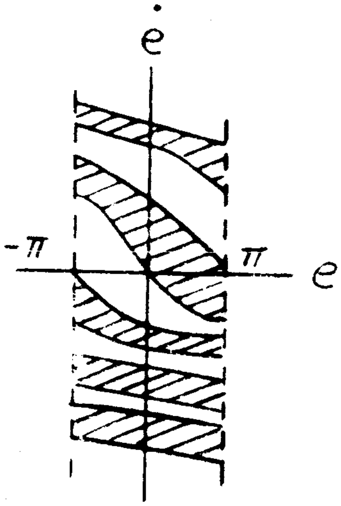


Figure 7.

Truncated Phase-plane Plot

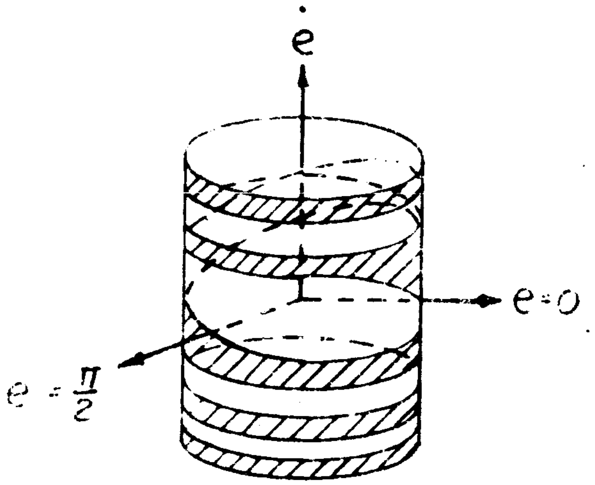


Figure 8.

Cylindrical Map of Phase-plane (ambiguity removed,.
(The shaded regions represent $o > 0$)

For example we note that the motor and load are in a closed system in which it is assumed

$$|T_d - fb| \leq F.$$

Clearly then $|\dot{e}| \leq \frac{2F}{f}$ is the worst case condition. If the e axis is expanded through choice of gear ratio then the semi-infite strip of Figure 9 can be represented in a finite box with as few as two discontinuities in traversing the range of e or $\dot{e}$. See Figure 10.

Another source of difficulty arises from the fact that no synthesis will be perfectly realized. Analogous to this are the well known oversimplifications of the signum operators,the inherent nonlinearities of the plant, and the inaccuracies of the sensors.

These realities lead one to believe that the system will have several torque transitions in its evolution to the origin. Usually this multiple switching is a limit cycle or a chattering regime or both.

We have, in our analog synthesis program, observed these phenomena in an attempt to put practical limitations on the acceptable tolerances of the servo components. Because our multiplier cabinets have been inoperative since the outset of this program, we have not as yet closed the loop on the analog synthesis of equation (13). We have, however, approximated equation (14) with straight line segments.

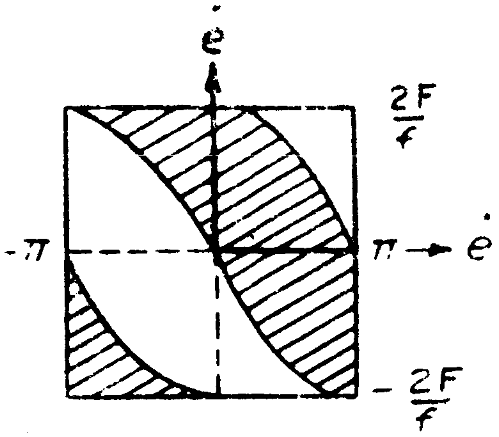


Figure 9.

Carthesian Plot of Phase-plane Showing High Gear Ratio.

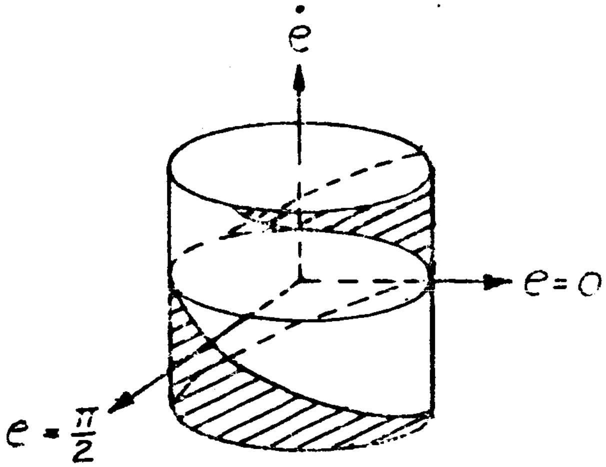


Figure 10.

Cylindrical Plot of Preceding Page Showing
High Gear Ratio.

List of Symbols used in 6.1

T = torque

τ_d = disturbance torque at motor shaft

T_d = disturbance torque at load

T_m = developed torque

T_L = load torque

J_L = inertia of load

J_m = inertia of armature

J = equivalent inertia at load

f_L = load friction

f_m = motor friction

f = equivalent friction at load

n = gear ratio

K_e = motor torque constant

K_n = motor damping constant

θ' = shaft displacement

θ = load displacement

E_{in} = terminal voltage

σ = switching function (or control function)

$F = \frac{K_e E_{in}}{n}$

$\theta_i(t)$ = input command

$e(t)$ = error
ϵ = base of natural logarithms
a = initial displacement of input command
$\text{sgn}\sigma$ = signum operator = 0 if $(\sigma = 0)$; $= \pm 1$ if $(\sigma \gtrless 0)$
b = rate of change of $\theta_i(t)$
t = time
e_o = initial error

$\dot{e}_o$ = initial error rate

A, B = undetermined coefficients of differential equations

6.2 Theory of Optimal Attitude Control and Stabilization of an Orbiting Vehicle.

For most purposes a satellite or space vehicle cannot be considered as a simple point particle. It must (at least) be considered as a rigid body, possessing three mutually perpendicular directions called "principle axes of inertia". The axes are unit vecotrs fixed in the body; their directions are determined by the body's geometry and mass distribution. The principle axes are taken as originating at the body's center of mass and as constituting a right-handed coordinate system. Specifically, the principal axes are unit vectors.

(1) $$u^1, u^2, u^3, \quad u^i = 1, \qquad (i=1,2,3)$$

(where $u^2 = u \cdot u$ and $\cdot$ denotes the scalar product) which are mutually perpendicular or orthogonal,

(2) $$u^i \cdot u^j = 0 \text{ if } i \neq j, \qquad (i,j=1,2,3).$$

We say that these axes constitute a body-fixed frame U,

(3) $$U = (u^1, u^2, u^3).$$

The frame U specifies the orientation of the body, relative to the inertial space E^3 which we now proceed to define. We assume that each vector u^i is specified by its components in an inertial frame I_3. An inertial frame is an orthogonal coordinate system fixed relative to the so-called "fixed stars"; the (non-relativistic) equations of motion of Newton and Euler are by definition valid in such a frame. Let the unit vectors e^1, e^2, e^3 be given by

(4) $$e^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad e^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad e^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} .$$

Then the frame

(5) $$I_3 = (e^1, e^2, e^3) = \text{diag}\ (1,1,1)$$

is specified by a square array of numbers, or a matrix, which in this case constitutes the identity matrix. We specify each vector u^i relative to I_3, i.e.

(6) $$u^i = u^i_1 e^1 + u^i_2 e^2 + u^i_3 e^3, \qquad (i=1,2,3)$$

where $u^i_j = e^j \cdot u^i$, $(j=1,2,3)$ are the components of u^i. Now we can regard (3) as a matrix equation, in which u^1, u^2, u^3 are the column vectors of the matrix U, and in which the results (1) and (2) are given more concisely by the statements

(7a) $$U^* = U^{-1}, \quad U^*U = UU^* = I_3,$$

where $*$ denotes matrix transposition (i.e. systematic replacement of the j^{th} column by the j^{th} row). The (equivalent) two relations of (7a) express the fact that the matrix U is orthogonal. Every orthogonal matrix preserves Euclidean length; that is, for any vector x in E^3,

(7b) $$\|Ux\| = \|x\| .$$

Indeed $\|Ux\|^2 = Ux \cdot Ux = x \cdot U^*Ux = x \cdot I_3 x = x \cdot x = \|x\|^2$; it can even be shown that if (7b) holds for every x, then (7a) holds. In fact, it should be clear by now that any orthogonal matrix defines the principal axis frame of some orientation in E^3. Hence we speak of orthogonal orientation matrices as defining the orientation, or attitude, of a rigid body in inertial space E^3 (which is given by the frame I_3).

The matrix U is specified by nine components u^i_j. However, by (7) there are six independent relations between these components, namely $\|u^1\| = \|u^2\| = \|u^3\| = 1$, $u^1 \cdot u^2 = u^2 \cdot u^3 = u^3 \cdot u^1 = 0$. This means that we should be able to specify U fully by at most $9 - 6 = 3$ independent parameters. In fact, it can be proved that for every orthogonal matrix U there corresponds a vector u in E^3,

$$u = (\theta_1, \theta_2, \theta_3)^* , \tag{8}$$

such that the columns of U, $u^i = u^i(u)$, $(i=1,2,3)$, are given by

$$u^1 = \begin{pmatrix} \cos\theta_2 \cos\theta_3 \\ \cos\theta_1 \sin\theta_3 + \sin\theta_1 \sin\theta_2 \cos\theta_3 \\ \sin\theta_1 \sin\theta_3 - \cos\theta_1 \sin\theta_2 \cos\theta_3 \end{pmatrix} , \tag{9_1}$$

$$u^2 = \begin{pmatrix} -\cos\theta_2 \sin\theta_3 \\ \cos\theta_1 \cos\theta_3 - \sin\theta_1 \sin\theta_2 \sin\theta_3 \\ \sin\theta_1 \cos\theta_3 + \cos\theta_1 \sin\theta_2 \sin\theta_3 \end{pmatrix} , \tag{9_2}$$

$$u^3 = \begin{pmatrix} \sin\theta_2 \\ -\sin\theta_1 \cos\theta_2 \\ \cos\theta_1 \cos\theta_2 \end{pmatrix} . \tag{9_3}$$

The geometrical interpretation of the (modified) Euler angles θ_1, θ_2, θ_3 is immediate:

$$U = U^1(\theta_1)\, U^2(\theta_2)\, U^3(\theta_3) \tag{10}$$

where

$$U^1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_1 & -\sin\theta_1 \\ 0 & \sin\theta_1 & \cos\theta_1 \end{pmatrix} , \tag{11_1}$$

$$U^2 = \begin{pmatrix} \cos\theta_2 & 0 & \sin\theta_2 \\ 0 & 1 & 0 \\ -\sin\theta_2 & 0 & \cos\theta_2 \end{pmatrix} ; \tag{11_2}$$

$$U^3 = \begin{pmatrix} \cos\theta_3 & -\sin\theta_3 & 0 \\ \sin\theta_3 & \cos\theta_3 & 0 \\ 0 & 0 & 1 \end{pmatrix} . \tag{11_3}$$

Thus, in order to move the frame I_3 into coincidence with the frame U, we proceed as follows. Firstly, rotate I_3 about the e^3 axis through an angle θ_3, obtaining the frame $U^3(\theta_3)$. Now rotate the frame U^3 about its second or u^2 axis through an angle θ_2, obtaining the frame U^2U^3. Finally, rotate the frame (U^2U^3) about its first or u^1 axis thrcugh and angle θ_1, obtaining the frame $U = U^1U^2U^3$.

Note that as $\|u\| \to 0$ the matrix $U = U(u) \to I_3$. In fact, by (3) and (9),

$$U(u) = I_\delta + \mathcal{O}(\|u\|^2), \tag{12}$$

where the notation $\mathcal{O}(\rho)$ means "$\leq \kappa\rho^2$ for all $\rho \leq \delta$, for some fixed numbers $\kappa > 0$, $\delta > 0$".

In aircraft work it is customary to take u^1 from the center of the plane to its nose, u^2 out the right wing and u^3 vertically downward. One then refers to the independant rotations $U^1(\theta_1), U^2(\theta_2), U^3(\theta_3)$ for small $\|u\| = (\theta_1^2 + \theta_2^2 + \theta_3^2)^{1/2}$, as *roll*, *pitch* and *yaw* maneuvers.

Now let $u = u(t)$ (hence $U = U(t)$) be specified as a function of time t, $0 \leq t \leq +\infty$. Let $\dot{} = d/dt$ denote the operation of taking rate of change with respect to time. Now we can define the *angular velocity vector* w of the frame $U = U(t)$ by

$$w = (1/2)(u^1 \times \dot{u}^1 + u^2 \times \dot{u}^2 + u^3 \times \dot{u}^3), \tag{13}$$

where $\otimes$ denotes the *vector product*

$$u \otimes v = \begin{pmatrix} u_2v_3 - u_3v_2 \\ u_3v_1 - u_1v_3 \\ u_1v_2 - u_2v_1 \end{pmatrix} . \tag{14}$$

(Note that in general $v \otimes v = 0$, while in (3), $u^3 = u^1 \otimes u^2$, $u^2 = u^3 \otimes u^1$, $u^1 = u^2 \otimes u^3$.) Using some elementary algebraic identities, it can be shown that (13) is equivalent to the *Poisson Equations* which refer to *rigid-body kinematics*:

$$\dot{u}^i = w \otimes u^i, \qquad (i=1,2,3). \tag{15}$$

For any vector $v = (v_1,v_2,v_3)*$, define the matrix

$$K(v) = \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix} . \tag{16}$$

Note that $K(v)$ is skew-symmetric, i.w. $K^* = -K$. Then it is easy to verify the identities

$$u \otimes v = K(u)v = -K(v)u, \quad K(v)v = 0, \tag{17}$$

whence (15) is equivalent to

$$\dot{U} = K(w)U, \quad U(0) = U_o. \tag{18}$$

The specification that the initial orientation of the body, U_o, and of the history of its angular velocity, $\{w(t) \mid 0 \leq t \leq +\infty\}$, is completely equivalent to the specification of the history of the orientation $\{U(t) \mid 0 \leq t \leq +\infty\}$. In fact, defining $A(t) = K(w(t))$, the orientation matrix U satisfies the differential equation

$$\dot{U} = A(t)U, \quad U(0) = U_o, \quad (A^*(t) = -A(t)). \tag{19}$$

It is easy to see that if U_o is orthogonal, then any solution $U(t)$ of (19) must be orthogonal.

In fact, $(U^*U) = \dot{U}^*U) + U^*\dot{U} = (AU)^*U + U^*AU = U^*A^*U + U^*AU = -U^*AU + U^*AU = 0$. Hence $U^*(t)U(t) = U^*_oU_o = I_3$. Furthermore, (19) can have at most one solution, for if $\tilde{U}$ and $\hat{U}$ both satisfy (19), then by linearity $\Delta U = \tilde{U} - \hat{U}$ satisfies (19), while $(\Delta U)_o = 0$. Thus

$$\Delta U(t) = \int_o^t A(\tau)\, \Delta U(\tau) d\tau. \tag{20}$$

Now define for any matrix M its norm $\|M\|$ as the smallest number such that

(21) $$\|Mx\| \leq \|M\| \, \|x\|$$

for every vector x. It can be shown that in general $\|AB\| \leq \|A\| \, \|B\|$, etc. Thus, defining $\varphi = \|\Delta U\|$, we have from (20),

(22) $$0 \leq \varphi(t) \leq \kappa + \int_0^t \alpha(\tau)\varphi(\tau)d\tau$$

where $\alpha(\tau)(= \|A(\tau)\|)$ is non-negative, and where $\kappa = \kappa < 0$ is any arbitrary number. (From (2), $\kappa = 0$, but we leave κ general in (22) for other reasons.) A fundamental lemma from differential equation theory asserts that, as a consequence of (22) it must be true that

(23) $$0 \leq \varphi(t) \leq \kappa \exp(\int_0^t \alpha(\tau)d\tau), \quad (0 \leq t \leq +\infty).$$

Since in (20), $\kappa = 0$, we have proved that $\Delta U(t) = 0$, i.e. that $U(t) = U(t)$ for $0 \leq t \leq +\infty$. Thus we have proved that (19) has at most one solution, once we have proved that (22) implies (23). As an illustration of Liapunov's Second Method we give a direct proof of (23). Define

(24) $$\psi(t) = (\kappa + \int_0^t \alpha(\tau)\varphi(\tau)d\tau) \exp(- \int_0^t \alpha(\tau)d\tau.$$

Note that

(25) $$\psi = -\alpha(t) \, (\kappa + \int_0^t \alpha(\tau)\varphi(\tau)d\tau) - \varphi(t) \, \exp(- \int_0^t \alpha(\tau)d\tau)$$

whence by (22), $\psi \leq 0$. Therefore $\psi(t) \leq \psi(0) = \kappa$, which is equivalent to (23). In conlusion, we shall prove that (19) has at least one solution. In fact, define the matrices $U^j(t)$, $(j=0,1,2,\ldots)$ by

$$(26) \qquad U^o(t) = U_o, \quad U^{j+1}(t) = U_o + \int_o^t A(\tau)U^j(\tau)d\tau.$$

Now for any fixed $T > 0$, there is (by continuity) a number $\kappa = \kappa(T)$ such that $\|A(t)\| < \kappa$ for $0 \leq t \leq T$. Hence (using $\|U^1 - U_o\| \leq \kappa t$), it is easy to prove by induction on j that

$$(27) \qquad \|U^{j+1}(t) - U^j(t)\| \leq (\kappa t)^j/j! \ , \qquad (j=0,1,2,\ldots)$$

for $0 \leq t \leq T$. Consequently there exists the limit

$$(28) \qquad U(t) = \lim_{j \to +\infty} U^j(t)$$

(for each fixed $t > 0$), and it is clear from (26) that $U = U_o + \int_o^t AUd\tau$, i.e., that U satisfies (19). In summary, we have proved that Poisson's Equation (19) possesses a unique orthogonal solution $U(t)$ for $0 \leq t \leq +\infty$.

We have gone into detail concerning integration of Poisson's Equation (19). This subject has a direct bearing on the manner in which the orientation of a rigid body relative to inertial space may be measured; and secondly, it is necessary to perform an at least approximate integration of Poisson's Equation on-board, in faster than real-time, if optimal control is to be achieved without use of a "closed-form" solution or a "pre-computed stored solution."

In"dead-reckoning" navigation, as practiced with great accuracy, for example, by Columbus, one computes one's present geographical positinn by means of a precise knowledge of one's initial position and a continuous record of one's speed and direction of motion at all subsequent times. In other words, if one's position be specified by the radius vector x in E^3, then one has for the kinematics of a point particle

(29) $$\dot{x} = v \quad (\dot{} = d/dt)$$

where v is the velocity vector of the particle; hence one's current position is computable from

(30) $$x(t) = x_o + \int_o^t v(\tau)d\tau.$$

In dead-reckoning navigation one simply performs an approximate numerical evaluation of the integral in (30).

Now in a space vehicle it is possible to measure the body's angular velocity vector $w(t)$ directly. In fact, three orthogonal rate gyros fixed along the body's principle axes will measure, respectively, w_1, w_2 and w_3. Furthermore, at some initial time the body's orientation U_o may be determined precisely by optical measurements involving the fixed stars. Consequently, an on-board computer capable of solving Poisson's Equation (19) in real time, where $A(t) = K(w(t))$, can provide a continuous estimate of the body's orientation, based on inertial sensors, between the times at which a more precise optical determination could be made.

With ground based computers, and arbitrary time available, Poisson's Equation can be integrated to any required numerical accuracy; in fact, to an accuracy such that only the erros in measurement of w affect the accuracy of the computed $U(t)$. The feasibility of such accuracy in real-time integration of (19), by an on-board computer, depends both on the merits of the numerical analysis used((26)-(27)-(28), though valid, is not particularly efficient or rapidly convergent) and on the state-of-the-art in computer technology. For many reasons the proposers regard the development of such an on-board computer as both desirable and inevitable.

Note that if (9) is inserted in (13), one finds that

$$(31) \qquad w = \begin{pmatrix} \dot{\theta}_1 + \dot{\theta}_3 \sin \theta_2 \\ \dot{\theta}_2 \cos \theta_1 - \dot{\theta}_3 \cos \theta_2 \sin \theta_1 \\ \dot{\theta}_2 \sin \theta_1 + \dot{\theta}_3 \cos \theta_2 \cos \theta_1 \end{pmatrix} .$$

Thus, if $\theta_2 \neq \pi/2$, and using $\dot{u} = (\dot{\theta}_1, \dot{\theta}_2, \dot{\theta}_3)$,

$$(32) \qquad \dot{u} = E(u)w, \quad (\theta_2 \neq \pi/2),$$

$$(33) \qquad E(u) = \begin{pmatrix} 1, \sin \theta_1 \tan \theta_2, - \cos \theta_1 \tan \theta_2 \\ 0, \cos \theta_1, \sin \theta_1 \\ 0, \sin \theta_1 \operatorname{cosec} \theta_2, \cos \theta_1 \operatorname{cosec} \theta_2 \end{pmatrix} .$$

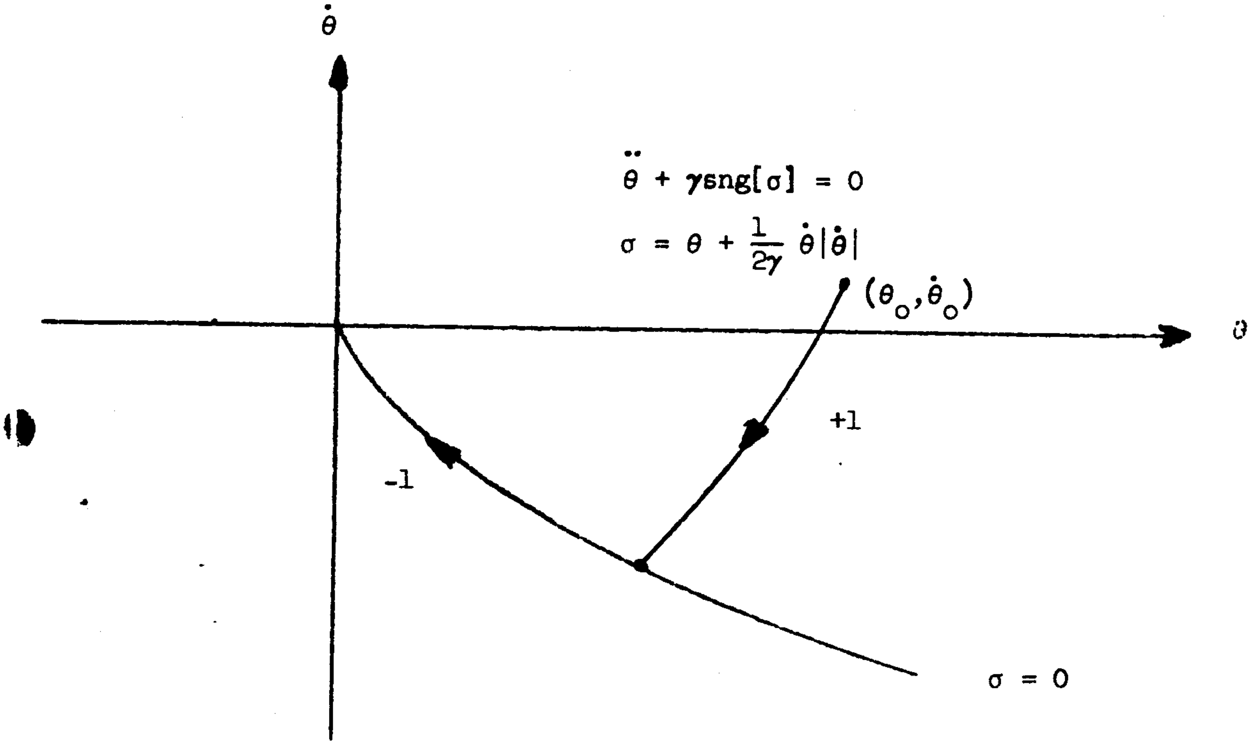
Clearly, $E(u) = I_3 + \mathcal{O}(\|u\|)$. Therefore,

$$(34) \qquad \dot{u} = w + \mathcal{O}(\|u\|\|w\|).$$

The question of whether use of Cayley-Klein's U or Euler's u is the best method for specifying a body's attitude is related to the question of integration of Poisson's Equation. Suppose that the angular velocity vector w is constant, i.e., that $w = w_o$. Then $\dot{U} = K(w_o)u$ can be integrated readily; in fact

$$(35) \qquad \dot{U} = K(w_o)U \iff U(t) = e^{K(w_o)t} U_o,$$

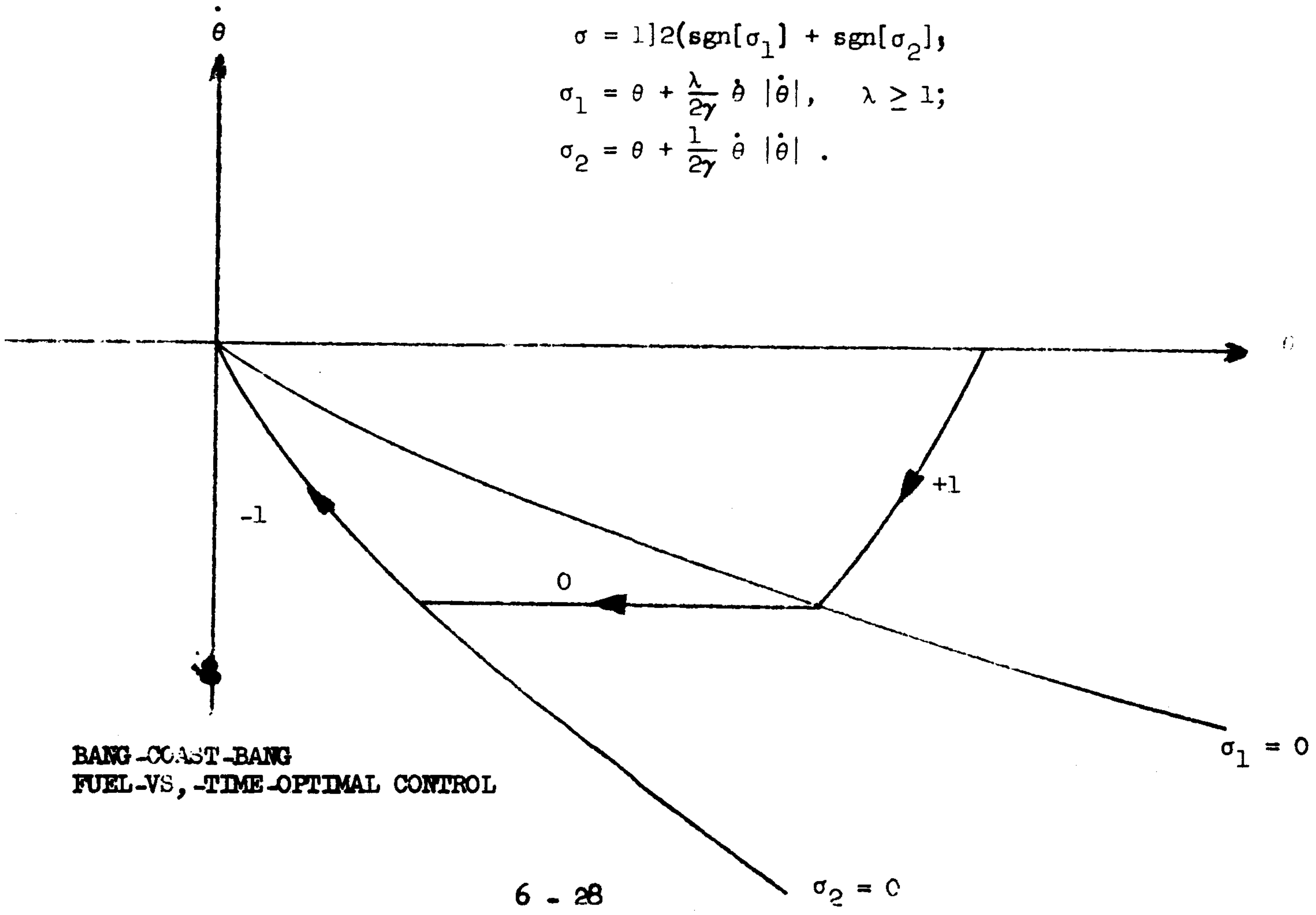
where, in general, we define



$$\ddot{\theta} + \gamma \mathrm{sng}[\sigma] = 0$$

$$\sigma = \theta + \frac{1}{2\gamma}\dot{\theta}|\dot{\theta}|$$

BANG-BANG TIME-OPTIMAL CONTROL



$$\sigma = 1]2(\mathrm{sgn}[\sigma_1] + \mathrm{sgn}[\sigma_2];$$

$$\sigma_1 = \theta + \frac{\lambda}{2\gamma}\dot{\theta}\ |\dot{\theta}|, \quad \lambda \geq 1;$$

$$\sigma_2 = \theta + \frac{1}{2\gamma}\dot{\theta}\ |\dot{\theta}|\ .$$

BANG-COAST-BANG
FUEL-VS,-TIME-OPTIMAL CONTROL

(36) $$e^{At} = I_n + \sum_{j=1}^{\infty} (t^j/j!)A^j \quad .$$

(it is easy to see that (36) converges for all t; in fact $\|e^{At}\| \leq e^{\|A\|t}$, $0 \leq t < +\infty$.)

The main drawback to the use of $U = (u^1, u^2, u^3)$ is its redundancy, caused by the fact that $U^*U = I_3$, i.e., that

(37) $$\|u^1\| = \|u^2\| = 1, \quad u^1 \cdot u^2 = 0, \quad u^3 = u^1 \times u^2 \; .$$

On the other hand, use of (35) contributes to a method of *monitoring reliability*, since we can check the accuracy of the computation of $e^{K(w_o(t))}$ by checking the validity of (37). In contrast, the use of the set of Euler angles u, while non-redundant, leads to a *nonlinear* Poisson Equation, even when w is a constant, namely

(38) $$\dot{u} = E(u)w_o, \quad u(0) = u_o.$$

It does not appear possible to find directly an explicit *closed form solution* $u(t) = f(t, u_o)$ to (38), although we can in theory invert the relationship $U = U(u)$ of (10) by means of an *implicitly* defined function $u = u(U)$, and then set $f(t, u_o) = u(e^{K(w_o)t} U(u_o))$. However, the actual discrepancy between (35) and (38) is not as great as might appear, since we can on the other hand find explicitly an independent set of *first integrals* of (38). Recall that a scalar function $\varphi(u)$ is a first integral of (38) if $\varphi(u(t)) = \varphi(u_o)$ for $0 \leq t < +\infty$. Similarly, a vector $h(u)$ is a *first integral vector* if

(39) $$h(u(t)) \equiv h(u_o), \quad (0 \leq t < +\infty).$$

Now we claim that

$$(40)\qquad \begin{pmatrix} w_o \cdot u^1(u) \\ w_o \cdot u^2(u) \\ w_o \cdot u^3(u) \end{pmatrix} = U^*(u)w_o$$

is a first integral vector of (38). In fact,

$$(41)\qquad d(h(u(t)))/dt = \dot{U}^*w_o = (K(w_o)U)^*w_o = U^*K(w_o)w_o \equiv 0.$$

Although each component of (39) is a first integral, they are not independent; in fact, obviously, $\|h(u)\| = \|U^*w_o\| = \|w_o\|$.

In this connection, note that $U(t) = e^{K(w_o)t}$ must be an orthogonal matrix by virtue of the skew-symmetry of K. In fact, clearly, $U^{-1}(t) = e^{-K(w_o)t}$ since $UU^{-1} = e^{K(w_o)(t-t)} = I_3$, while $U^* = e^{K^*(w_o)t} = e^{-K(w_o)t} = U^{-1}$. This fact suggests the approximate integration of Poisson's Equation (18) by the following piecewise constant or step-function approximation. For a very small sampling period τ, define

$$(42a)\qquad U(t) \cong U^j, \quad j\tau \leq t \leq (j+1)\tau,$$

$$(42b)\qquad U^{j+1} = e^{K(w(j\tau))\tau}U^j, \quad (j = 0,1,2, \dots).$$

Whether or note $w(j\tau)$ is measured with perfect accuracy, K is perfectly skew-symmetric by construction, whence each U^j is automatically orthogonal to the extent that $e^{K\tau}$ has been computed accurately and to the extent that the accumulation of round-off errors in passing from U^j to U^{j+1} has not become serious. For very small τ, the series (36) can be successfully truncated after the first term, which leads to

$$(42c) \qquad U^{j+1} = [I_3 + \tau K(w(j\tau))]U^j + \quad (\tau^2), \quad (j = 0,1,2, \ldots).$$

In fact, such a U^{j+1} is approximately orthogonal because $(I_3 + \tau K)^* = (I_3 - \tau K) = (I_3 + \tau K)^{-1} + \quad (\tau^2)$, by virtue of the C. Neuman resolvent series

$$(43) \qquad (I_3 + \tau K)^{-1} = I_3 + \sum_{j=1}^{\infty} (-1)^j (\tau K)^j$$

which converges for all $|\tau| < i/\|K\|$.

The proposers feel that a more careful examination of the propagation of truncation and round-off errors in the numerical scheme (42), in connection with the state of the instrument art (threshold, drift rates) in w-sensors and certain new computer organization (cf. Aeronca's "parallel systems" or "relaxation" computer) will establish the feasibility of a new type of on-board computer for real-time integration of the Poisson Equation.

Turning now to the dynamics of rigid bodies in space, the virtues of the Euler frame U will be manifest. Let g_i denote the total external torque applied to the body's u^i axis $(i = 1,2,3)$, and let $g = (g_1, g_2, g_3)^*$ denote the total torque vector. (Note: g is expressed now in the U frame; to express g in the inertial space E^3, we must use the vector $g = U^*g = U^{-1}g$.) Next, let J_i be the moment of inertia of the body computed about the u^i axis, $(i = 1,2,3,)$, and note that in the U frame the body's inertia tensor can be represented by the matrix

$$J = \operatorname{diag}(J_1, J_2, J_3) = (J_1 e^1, \ J_2 e^2, J_3 e^3).$$

Now, if J is constant, the evolution in time of the angular velocity vector w is governed by Euler's Equation, which together with Poisson's Equation is

(44a) $$\dot{U} = K(w)U,\ U(0) = U_o; \quad \text{or} \quad \dot{u} = E(u)w,\ u(0) = u_o,\ U = u(u);$$

(44b) $$J\dot{w} + K(w)Jw = g, \quad w(0) = w_o, \quad (\dot{J} = 0).$$

If $g = g(t)$ is specified only relative to the inertial space E^3 by $g(t)$, then $g = Ug(t)$, and we must adjoin to (44b) the Poisson Equation. However, if $g(t)$ is defined only relative to the body, then (44) can be integrated independently of Poisson's Equation.

For example, if $g = g(w)$ has the property that

(45) $$w \cdot g(w) < 0, \quad (w \neq 0)$$

then $w(t) \to 0$ as $t \to +\infty$. In fact, using Liapunov's Second Method, let

(46a) $$\varphi = (1/2)w \cdot Jw$$

and compute that

(46b) $$\dot{\varphi} = w \cdot J\dot{w} = -w \cdot K(w)Jw + w \cdot g < 0, \quad (w \neq 0),$$

since

$$w \cdot K(w)Jw = K^*(w)w \cdot Jw = -K(w)w \cdot Jw \quad \text{and} \quad K(w)w \equiv 0.$$

A similar effect can be obtained by altering the geometry of the body's mass distribution, in which case J is not a constant.

In one satellite (TIROS) weights were released, on cords, which moved away because of centrifugal acceleration. However, as the weights moved away, increasing the body's radius, J_o increased, which decreased w_o. In theory one could obtain $w_o \to 0$ by allowing $J_o \to +\infty$. In practice, this was done by releasing the weights after w_o was reduced to an acceptable level. Another similar scheme seriuusly proposed for attitude control involves extension or retraction of lengthy telescopic booms made of wide rolls of thin but stiff plastic. Such schemes can, however, change a vehicle's attitude or angular velocity only if $w_o \neq 0$.

For this reason it is often acceptable to regard J as a constant, treating the effects of $\dot{J} \neq 0$ by replacing g by $g-\dot{J}w$, i.e., by regarding the term $-\dot{J}w$ as an external disturbing torque whose effects must be overcome by proper disposition of the control torque g. If J is changed deliberately but rather occurs randomly (e.g., pilots moving about in a manned spacecraft) then this is doubtless the correct method of treatment of the torque $-\dot{J}w$.

While on the subject of external perturbing torques, it should be noted that for a high-altitude satellite these torques are quite small. They include

(i) residual atmospheric drag;

(ii) meteoric dust impacts;

(iii) gravitational gradient torques;

(iv) radiation pressure from sun;

(v) magnetic field interactions and induced electric charges.

In the case of the Transit satellites, permanent bar magnets of exceptionally powerful gaussian strenght have been actually used to damp $\dot{w}$ to zero precisely as in (45) and (46). More generally, the interaction of torques (iii) and (v) have led to various phenomena observed experimentally in the attitude histories of one Explorer satellite and in a Tiros satellite. (The satellites possessed magnetic field both by virtue of residual permanent magnetism and by circulating currents in their payloads). Here we have the phenomenon of an external torque which cannot be specified except in relation to the surrounding inertial space. Specifically the earth's gravitational and magnetic fields can be regarded as having known histories and futures in E^3; even in a first approximation, regarding the earth as a fixed object in E^3 and the satellite in a known almost periodic orbit (including precessional effects due to the earth's oblateness and consequent non-spherically-symmetric gravitational field acting on the satellite considered as a point mass); we must represent g as of the form

$$g = g(t, U) \tag{47}$$

where $g(t, U)$ is almost periodic in t, but where we cannot consider Euler's Equation separately from Poisson's Equation. The Explorer satellite experienced an almost-periodic large fluctuation in its angular velocity w which is still regarded as a mystery by the cognizant NASA physicists. The Tiros exhibited a very large fluctuation both in w and U (the ranges of θ_1, θ_2, θ_3 exceeding 90°) with an almost-period of many days; subsequently NASA and RCA scientists performed a numerical integration of (44) with a suitable term (47) based on torques (iii) and (v) and obtained close agreement with the previously recorded observations. There is still not available, however, an adequate analytical theory of this subject, nor a method for treating it other than by numerical integration of (44). Nevertheless, it has been seriously proposed that the attitude of a satellite can be controlled (in particular, by radio-relay from a ground-based control-computer) by deliberate variation in circulating current loops within the satellite. Hitherto no scientific approach to the synthesis of such a system has been suggested; however, it will become apparent that the techniques discussed below by the proposers are sufficiently comprehensive as to include even such a subtle synthesis problem as this one.

From the point of view of control synthesis, where the actuating torques are to be independent of the phenomena (i)-(v) above, and independent of $\dot{J}$, it seems best to lump all of the torques into a single torque regarded as an unknown forcing term or "random input" $d(t)$. Thus the idealized equations of attitude control are

$$(48a) \qquad \overset{\circ}{U} = K(w)U, \quad U(0) = U_o; \quad \text{or} \quad U = U(u), \ \dot{u} = E(u)w, \ u(0) = u_o$$

$$(48b) \qquad J\dot{w} + K(w)Jw = g + d(t), \quad w(0) = w_o.$$

We can assume that many properties of $d(t)$, in particular $\max\|d(t)\|$, can be predicted from statistical studies of the environment; however, $d(t)$ is to be regarded as unknown otherwise. The idealized attitude control problem can be stated as follows. Given a desired terminal state (U^1, w^1), find a control law g such that, from the given state (U_o, v_o) the state $(U(t), w(t))$ evolves for $0 \leq t \leq T$ until

$$(49) \qquad (u(t), w(t)) \to (U^1, w^1) \quad \text{as} \quad t \to T \quad (T \leq +\infty).$$

There are many variations on this theme. If T is a fixed a priori we speak of terminal control. It may be impossible, for small T, unless we pay the cost of a sifficiently large control torque g. In scientific satellites, T need not be small; in fact, it may be minutes, hours or even days. In future manned spacecraft, especially in military operations (or in automatic orbiting anti-ICBM satellites) the transition time T can be critical; hence we expect the importance of time-optimal control to increase. At any rate, in non-terminal control, the time T is not given in advance but determined implicitly by (49). If g is continuous and bounded, then necessarily $T = +\infty$. If g is allowed to be discontinuous (as, e.g., with reaction jets switched on and off) then we can take T to be finite.

If $g = g(t; w_o, U_o)$ then we speak of pre-programmed or open-loop control. If one measures the state (w, U) continuously and if $g = g(w, U; w^1, U^1)$ then one speaks of feedback or closed-loop control. The virtue of the latter is that, when $\Delta w = w - w^1$ and

$\Delta U = U-U^1$ are small, most control laws g can be expressed as $g = g(\Delta w,\Delta U)$, whence it is not necessary to use a different law g for each state (w^1, U^1) whose acquisition is desired. Furthermore, if $g(0,0) = 0$ but $g \neq 0$ for $\|\Delta w\|^2 + \|\Delta U\|^2 \neq 0$, then a feedback control operates until the measured acquisition error is zero, which in view of the imperfections in all measurement, computations and mechanizations, is epistemologically preferrable; for then the only thing certain is that zero error will be attained at least within the threshold limits of the sensors used to establish w and U. However, feedback control, though more accurate, is also more expensive, more complex and hence less reliable than preprogrammed control.

We can also consider the efficiency of the control law chosen. Normally, various *a priori* constraints on g are available such as

(50) $$|g_i| \leq \gamma_i \quad (\gamma_i \geq 0; \quad i = 1,2,3).$$

We can define as a performance criterion, or basis of comparison of possible control laws satisfying the constraints, an integral such as

(51) $$\pi = \pi(w_o, u_o, g) = \int_0^T \alpha(w,U,g)dt$$

Here $\alpha = \alpha(w,U,g) \geq 0$ is a non-negative smooth function. One can then define an admissable $\{g\}$ as optimal relative to π if it minimizes π in comparison with other admissable $\{g\}$'s .

In Aeronca's Air Force Contract AF 33(616)8285, Monthly Progress Report No. 6, page 35, the following result is proved.

THEOREM. For any admissable performance criterion, the rigid-body system (47) can be optimally controlled.

Now if $\{g\}$ is admissable and π is well defined for all (w_o, u_o) in some domain, then we may define $x = (u^*, w^*)^* = \binom{u}{w}$, and so define $\pi = \pi(x_o; \{g\})$ for all x_o in . Then we can define a vector

(52) $$p = p(x) = -\mathrm{grad}_{(x)}\pi$$

for each $\{g\}$. Now we can write the attitde control system as

(53) $$\dot{x} = (J^{-1}E(u)w[-K(w)Jw + g(x)]) = F(x,g(x)),\ x_o \text{ in } .$$

The content of the statement (51) is that $\pi = \pi(x)$ is a Liapunov function for (53). In fact, by (51)

(54) $$d\pi(x(t))/dt = -\alpha(x) < 0.$$

If $\alpha(0) = 0$, then $T = +\infty$. If $\alpha(x) \geq \delta > 0$, then $\pi(x(T)) = 0$, i.e., $x = 0$, for some $T \leq T_o/\delta$.

Thus we define

(55) $$H = H(x,p,g) = p \cdot F(x,g) - \alpha(x;\ g)$$

The equivalent statements (51) and (54) are also now obviously equivalent to the statement that

(56) $$H = 0.$$

In other words, a control law g controls the system (53) and defines a performance index π, if and only if the associated Hamiltonian is zero.

Furthermore, among all admissable control laws $\{g\}$, i.e., laws for which $H = 0$, that one (or ones) is optimal which *maximizes* H relative to the constraints such as (50); in fact, *if the control law* g *is such that* $H(x,p,g) \equiv 0$, *then* g *is also optimal if and only if the associated Hamiltonian is maximal, i.e.*,

$$H(x,p,g) = \overline{H} = \max_{|g_i| \leq \gamma_i} H(x,p,g). \tag{57}$$

In practice, (57) determines the optimal g as a function of x and $p = -\text{grad}_{(x)}\pi$. Let us call this function $\bar{g}(x;\, p)$. Inserting this into the statement $H = 0$, we have a partial differential equation

$$H(x,p,\bar{g}(x,\ -\text{grad}\ \pi)) = 0, \tag{58}$$

the *Hamiltonian-Jacobi* equation which determines π. Explicitly,

$$f(x,\bar{g}(x,\ -\text{grad}\ \pi))\cdot \text{grad}\ \pi = -\alpha(x,\bar{g}(x,\ -\text{grad}\ \pi)). \tag{59}$$

Although this equation appears to be formidable, the constraints $|g_i| \leq \gamma_i$ usually require that g be piecewise constant, i.e., that $|g_i| = \gamma_i$. (However, if $\alpha(x,\ g)$ is quadratic in both x and g ,i.e, $\alpha(x,g) = x\cdot Cx + g\cdot Qg$, then in certain regions g will be linear in x.) In this case we can piece toegether $\pi(x)$ from solutions of the equation'

$$F(x,k)\cdot \text{grad}\ \pi = -\alpha(x,k) \tag{60}$$

where k is an arbitrary constant such that $|k_i| = \gamma_i$. The proposers have completely, in all details, solved the problem (59) for control of linear plants $(F = F^0x + F^1g)$. For the case of symmetric satellites $J_1 = J_2$ they have solved the preliminary problem (60) and hope, in due time, to correctly piece together the known functions $\pi(x,k)$ into the global definition of $\pi(x)$ on , in which case there will be available a closed form solution for the synthesis of optimal satellite attitude control systems.

Another method for computing the optimal $g(x)$ is based on the recent discovery that if one defines a family of control laws $g = g(x;\mu)$ by

$$(61) \qquad \frac{dg}{d\mu} = -\text{grad}_{(g)} H(x,p(x),g), \quad (0 \leqq \mu < + \infty)$$

then

$$(62) \qquad \frac{d\pi}{d\mu} = - \int_0^T \|\text{grad}_{(g)} H\|^2 dt < 0.$$

Hence as $\mu \to + \infty$, π decreases to its minimum and $g(x; \mu)$ evolves to the optimal $g(x)$. This steepest descent computation is suitable for a faster-than-real-time on-board computer such as Aeronca's Relaxation Computer, which for each measured state x almost instantly computes the optimal $g(x)$ by integrating (61) with an arbitrary initial control law $g(x; 0)$.

If one wishes to obtain truly high performance, one should take the control law $g = g(u,w; d)$ to be a function of the random disturbing input d, which is not known and exceedingly difficult to measure directly. However, a time-optimal self-adaptive servomotor giving great gains in performance can be obtained by regarding d (which for our Optimotor is the input rate, and for the Saturn space vehicle is the crosswind velocity) as piecewise constant and approximating it by indirect measurements utilizing more readily observed variables For example, if we regard g and w as observable, and $d = d(\theta)$ constant for $t-\tau \leqq \theta \leqq t$, then from (48)

(63a)
$$d = \frac{1}{\tau} \{J(w(t)-w(t-\tau)) + \int_{t-\tau}^{t} K(w(\theta))Jw(\theta)d\theta - \int_{t-\tau}^{t} g(\theta)d\theta\}$$

$$= \int_{t-\tau}^{t} h(w(\theta), g(\theta))d\theta.$$

In particular, if we use a closed form

(63b)
$$g = g(w,u,d)$$

for optimal control of (48), then the control law

(63c)
$$g = g\{w(t), u(t), \int_{t-\tau}^{t} h(w(\theta), g(\theta))d\theta\}$$

will provide an optimal nonlinear feedback law rendered optimally self-adaptive to random disturbances by means of nonlinear integral feedback. Notice that for a sufficient large integer N,

(64)
$$\int_{t-\tau}^{t} h(\theta)d\theta = \frac{\tau}{N} \sum_{j=0}^{N-1} h(t - \frac{j\tau}{(N-1)}) + O((\tau/N)^2)$$

which indicates that one can mechanize (64) quite readily by means of a sampled-data control computer with a memory capacity for retaining the system state for N sampling periods in the past.

A detailed example of (63b) and (63c) will be presented below in equations (82)-(84).

Consider now the mechanization of the desired control law g. The torque g is to be produced by actuators, and the dynamics of the actuators must be considered. Of course, the simplest method is to use pairs of reaction jets, in which case we may write

(65) $$g = G \operatorname{sgn}[c] \quad G = \operatorname{diag}(\gamma_1, \gamma_2, \gamma_3)$$

(where by $\operatorname{sgn}[c]$ we mean $(\operatorname{sgn}[c]) \circ e^i = e^i \circ c / |e^i \circ c|$, $(i = 1,2,3)$). The major consideration regarding the use of (65) is that the phenomena of time-delay dead-zone and hysteresis (present in any physical relay or switch) lead to a control system in which, after nearly attaining the desired state, the system's state oscillates or "hunts" around it in a small stable limit-cycle. (Even if a dead-zone is employed, the limit-cycle always exists when the disturbing torque $d \neq 0$).

Several effective methods for analytical study of the amplitude and frequency of the limit cycle as a function of the characteristics of the actuators and sensors are given in Aeronca Technical Reports Nos. 60-14 and 60-16.

Consider finally the question of sensing the satellite's state (U, w). Even if we know precisely the desired control law $g = g(u,w)$, we must mechanize it by using in (48), not this g but rather

(66) $$g = g(U_*, w_*)$$

where (U_*, w_*) constitutes an approximate measurement of the state. In the past it has been customary to assume that the sensing instruments obey linear laws, e.g., that

(67) $$\dot{U}_* = K(w_*)U_*, \quad \ddot{w}_* + R\dot{w}_* + Cw_* = Gw$$

where R and G are diagonal matrixes of positive elements. However, for the acquisition problem this is not adequate as we have proved quite rigorously.

In measuring a physical macroscopic quantity, something in a way similar to what is described by the Uncertainty Principle of Microphysics takes place. In fact, when the quantity being measured is relatively small, the measuring instrument alters that quantity (and eventually other dynamical variables of the system) to a degree which cannot be reduced below certain natural practical limits. A detailed examination of the dynamics of satellite-borne inertial instruments establishes this fact quantitatively (see Aeronca Final Report on "Optimal-Nonlinear Systems for the Attitude Control of an Orbiting Vehicle," contract AF 33(616)8285).

Regarding selection of performance criteria, we say that while time-optimality is basic to many military missions and should not be neglected, criteria regarding fuel-mass expenditure or energy cost may be paramount in the average acquisition maneuver. However, it makes no sense to consider either of the latter except as a trade-off against time, for otherwise an infinitely slow maneuver uses the least fuel or energy.

In the use of jet reaction control

(68) $$\dot{u} = E(u)w, \quad J\dot{w} = -K(w)Jw + Gc + d$$

we shall employ either

(69) $$\pi_1 = \int_0^T (1 + \mu|c|)dt, \qquad \mu > 0,$$

where $|c| = c \cdot \mathrm{sgn}[c] = |c_1| + |c_2| + |c_3|$, in which case we obtain a discontinuous control $c = -\mathrm{sgn}[p]$ which minimizes $\pi_1 = (\text{transition time}) + \mu(\text{fuel-mass})$;

(0) $$\pi_2 = \int_0^\infty [\lambda(w \cdot Jw) + \mu(c \circ Gc)]dt, \quad \lambda > 0, \quad \mu > 0,$$

in which case we find a continuous control law, necessitating throttling of the jets for minimization of (3).

An important result in this connection is the superposition Principle for Euler's angles. In fact, if in the preceding we choose $(u^{1,1}, \; u^{2,1}, \; u^{3,1}) = (e^1, e^2, e^3) = I_3$, then we obtain a standard control law.

(71) $$c = -\mathrm{sgn}[g], \quad g = g(\theta_1, \theta_2, \theta_3, w_1, w_2, w_3)\ .$$

The control law (71) drives the rigid-body from any initial state to rest in the standard orientation $u(T) = w(T) = 0$, i.e.,

(72) $$\theta_1(T) = \theta_2(T) = \theta_3(T) = w_1(T) = w_2(T) = w_3(T) = 0.$$

It is of considerable importance that the control law

(73) $$g = g(\theta_1 - \theta_1^{\,1},\ \theta_2 - \theta_2^{\,1} \cdot \theta_3 - \theta_3^{\,1},\ w_1, w_2, w_3)$$

drives the body optimally to rest in the orientation

(74) $$\theta_1(T) = \theta_1^{\,1},\ \theta_2(T) = \theta_2^{\,2},\ \theta_3(T) = \theta_3^{\,3},\ w(T) = 0.$$

The availability of the <u>superposition principle</u> (73)-(74) is quite convenient; however, there is no such principle available if we abandon the condition that $w(T) = 0$. Thus, optimal control to a slewing state $w(T) = w^1$, cannot be obtained merely by using $g = g(U-U^1, w-w^1)$; rather the control law has the form $g(U-U^1, w, w^1)$.

The second possible computrol system relies on solving the <u>two-point boundary value problem</u> with an on-board high-speed computer. Studies indicate that of the numerous schemes proposed for this the most efficient is the Relaxation Method described below in Section

The third method is the <u>closed form solution</u> technique for optimal control system synthesis. In the sequel, we shall derive the linear and quadratic terms in the power series expansion of the optimal control law and prove that this <u>Approximate Closed Form</u> yields a stable control system which is quasi-optimal. A preliminary design of Aeronca's <u>Closed Form Computer</u> will also be presented.

For time-optimal control each jet-actuated control torque should either be given by the signum of + 1 or -1. Such a control is called <u>BANG-BANG control</u>.

In 2.3.2 a discussion of BANG-COAST-BANG control fuel-mass-minimal-control with a concrete example from attitude control is given. From this discussion a control law stable as a simultaneous 3-axis quasi-optimal control law in some neighborhood of the origin is developed. In practice we may, for example, attain this neighborhood by using the Stored Function Computer, then switch to the Relaxation Computer or the Closed Form Computer, for bringing the body to the state wherein the linear vernier control is to apply.

7.0 Conclusions and Recommendations.

It is now clear that the theory of optimal control is able to yield closed form expressions for the switching surfaces. That is, the theory has given us the condition of the driving of the actuators and from this condition it is possible to state the switching surfaces for an admittedly small class of plants. Further while this class of plants is small, it is general in the sense that the phase space can be of arbitrarily high dimension.

Now examine what is meant by Optimal Control. By this is meant a control that will give the best performance with respect to some constraint. That expressions for the switching surfaces are available means that it is possible to know the procedure that must be followed to obtain the ultimate performance, but more so it is possible to have insight as to how to best approach this performance in view of more general constraints that arise with a real plant which are not so easily stated in anlitic terms. Further it is possible to have some criteria of performance in the selection of a feasible control system, where it is not possible to make the feasible control optimal. In essence it is now possible to begin the development of a "rational" synthesis procedure.

Above in Chapter 3, areas of attack have been suggested. These procedures were suggested from the standpoint of simply pushing the problem of constructing the control surfaces. Here the more general question of developing a "rational" synthesis procedure will be examined.

First there is of course the problem of pushing the theory toward developing control surfaces for more general plants. But there is another direction in which the theory can be enlarged. Now that expressions for the surface are available, it is desirable to know what can be done with them. In particular, it might arise that in the control of a nonlinear plant that one would want to use,the surface that would arise from locally linearizing the abstract plant for every point in phase space. Would this give a control law and if so under what conditions for the plant. Further if such a procedure did make the plant controllable, by what performance criteria could it be determined how good the control is.

Another of the observations that is to be made about the results now in hand is how unwieldly they become for even the cases of small order. On the other hand, it seems that there will be methodical procedures for generating the surfaces. It seems that a situation is arising that is not amenable to ordinary analytical techniques. A step toward more abstract methods must of necessity be made if the information required is to be elicited. The designer is not going to have statements at hand in a form simple enough so that he will be able to immediately go to a computer and generate the data with which to make his decision. Rather it looks like he is going to have to go to the computer to even generate the statements that he wishes to examine, and subsequently have to use the computer to scan the statements to deliver the criteria he seeks. Prior to the possibility of this, much more about the abstract qualities of the expressions for the switching surfaces must be known.

The recommendations are then just an expression of the above comments.

I. That work be continued toward the development of expressions for switching surfaces for more general plants.

II. That techniques for generating and analyzing these expression by the use of computers be developed as an adjunct to recommentation I. and for the purpose of developing a "rational"synthesis procedure.

III. That the design of a specific plant or plants be carried out as a focus for learning how to integrate the theory and techniques now known.

APPENDIX I

The polynomial

(1) $$x^r + s_1 x^{r-1} + \cdots + s_r = 0$$

has the companion matrix

(2) $$\begin{vmatrix} -s_1 & 1 & 0 & \cdots & \\ -s_2 & 0 & 1 & & \\ \cdots & & & & 1 \\ -s_r & \cdots & & & 0 \end{vmatrix} = S$$

Each eigenvalue λ of the matrix satisfies the relation

$$SZ = \lambda Z$$

for the eigenvector z.

Specifically, for the eigenvector having first term $= 1$,

(3) $$\lambda = \frac{z_{i+1} - s_i}{z_i} \quad \text{for all } i$$

and let

$$n_i = z_{i+1} - s_i .$$

We let the consequence (3) become our condition and define for arbitrary Z not necessarily an eigenvector the set of λ's,

(4) $$\lambda_i = \frac{z_{i+1} - s_i}{z_i} = \frac{n_i}{z_i} .$$

Make the numbers

(5) $$\lambda_i - \lambda_j = \frac{n_i}{z_i} - \frac{n_j}{z_j} .$$

Define

(6) $$u_{ij} = n_i z_j - n_j z_i$$

It is to be noted that an if and only if relationship exists between (5) and (6) for non zero z_i's. More pertinent is that the condition of z being an eigenvector makes all u_{ij} equal to zero. By the same token the function

(7) $$\psi = \sum_{ij} u_{ij}\bar{u}_{ij} ,$$

will also be zero under this condition. As a sum of moduli it is obvious that ψ can only take values $\geq$ than zero at points other than eigenvectors.

The one remaining possibility that ψ could be zero is for

(8) $$Z = 0$$

which is eliminated by constraining z_1 to be equal to 1. Finding the points at which $\psi = 0$ then finds the eigenvectors of the matrix S.

The procedure that is used is that of steepest descents. An arbitrary value of Z is chosen. The gradient of ψ for the arbitrary Z computed and then the best distance along the gradient is determined. This pair,the gradient and the best distance, give a correction to Z and the process is then repeated until a final converged value of Z is obtained.

The algebra of this procedure follows. First the gradient

(9) $$\nabla \psi = \frac{\partial\psi}{\partial Z}$$

Variation of the ψ and gathering the coefficients of the varied components of Z, δz_i gives these coefficients.

(10) $$\delta\psi = 2R\Big(\sum_{ij} (\delta z_i n_j + \delta n_j z_i - \delta z_j n_i - \delta n_i z_j)\bar{u}_{ij}\Big) = 2R(2\gamma)$$

(11) $$\begin{aligned} 2\gamma = \sum_{ij}(&\delta z_i n_j \bar{z}_i \bar{n}_j + \delta n_j z_i \bar{z}_i \bar{n}_j - \delta z_j n_i \bar{z}_i \bar{n}_j \\ &- \delta n_i z_j \bar{z}_i \bar{n}_j - \delta z_i n_j \bar{z}_j \bar{n}_i - \delta n_j z_i \bar{z}_j \bar{n}_i \\ &+ \delta z_j n_i \bar{z}_j \bar{n}_i + \delta n_i z_j \bar{z}_j \bar{n}_i) \end{aligned}$$

(12) $$k = \sum_i |z_i|^2$$

$$\ell = \sum_i |n_i|^2$$

$$A = \sum_i n_i \bar{z}_i$$

(13) $$\gamma = k\sum_i \bar{n}_i \delta n_i + \ell \sum_i \bar{z}_i \delta z_i - A\sum_i \bar{n}_i \delta z_i - \bar{A}\sum_i \bar{z}_i \delta n_i$$

(14) $$\frac{1}{2}\,\delta_i\psi = R(k\bar{n}_{i-1} - A\bar{n}_i + \ell\bar{z}_i - \bar{A}z_{i-1})\delta z_i$$

(15) $$\nabla\psi = kn^{-} + \ell z - An - \bar{A}z^{-}$$

Improving Z by the vector $R'\bar{v}$ in the direction $V = \nabla\psi$, gives the polynomial

(16) $$u_{ij} = (z_i + r'v_i)(n_j + r'v_{j+1}) - (z_j + r'v_j)(n_j + r'v_{i+1})$$

(17) $$u_{ij} = \alpha_{ij}\, r'^2 + \beta_{ij} r' + \gamma_{ij}$$

where

$$\alpha_{ij} = (v_i v_{j+1} - v_j v_{i+1})$$

$$\beta_{ij} = (z_i v_{j+1} - z_j v_{i+1} + n_j v_i - n_i v_j)$$

$$\gamma_{ij} = (z_i n_j - z_j n_i)$$

and for ψ

(18) $$\psi = \sum_{ij} (\alpha_j r'^2 + \beta_{ij} r' + \gamma_{ij})(\bar{\alpha}_{ij} \bar{r}' + \bar{\beta}_{ij} \bar{r}' + \bar{\gamma}_{ij})$$

where

(19) $$|r'|^4 \sum_{ij} \alpha_{ij} \bar{\alpha}_{ij} = |r'|^4 (c - |B|^2) = a_o |r'|^4$$

$$|r'|^2\, r' \sum_{ij} \alpha_{ij} \bar{\beta}_{ij} = |r'|^2 r' (c(c + E) - BD - \bar{B}F) = \bar{a}_1 |r'|^2 r'$$

$$|r'|^2 \bar{r}' \sum_{ij} \bar{\alpha}_{ij} \beta_{ij} = a_1 \bar{r}' |r'|^2$$

$$|r'|^2 \sum_{ij} \beta_{ij} \bar{\beta}_{ij} = |r'|^2 (c(a + b) + E\bar{C} + \bar{E}C - AB - \overline{AB} - |D|^2 - |F|^2) = a_2 |r'|^2$$

$$r'^2 \sum_{ij} \alpha_{ij} \bar{\gamma}_{ij} = r'^2 (CE - DF) = r' \bar{a}_3$$

$$\bar{r}'^2 \sum_{ij} \bar{\alpha}_{ij} \gamma_{ij} = \bar{r}'^2 a_3$$

$$r'\sum_{ij}\beta_{ij}\bar{\gamma}_{ij} = r'(aE - D\bar{A}) = r'\bar{a}_4$$

$$\bar{r}'\sum_{ij}\bar{\beta}_{ij}\gamma_{ij} = \bar{r}'a_4$$

$$\sum_{ij}\gamma_{ij}\bar{\gamma}_{ij} = ab-|A|^2 = a_5$$

and

$$k = \sum_i |z_i|^2$$

$$\ell = \sum_i |n_i|^2$$

$$m = \sum_i |v_i|^2$$

$$A = \sum_i n_i\bar{z}_i$$

$$B = \sum_i v_i\bar{v}_{i+1}$$

$$C = \sum_i v_i\bar{z}_i$$

$$D = \sum_i v_{i+1}\bar{z}_i$$

$$E = \sum_i v_{i+1}\bar{n}_i$$

$$F = \sum_i v_i\bar{n}_i$$

It is to be noted that the equation is a polynomial in both r' and $\bar{r}'$. As $\bar{r}'$ is not an analytic function of r' a straightforward solution is not available. In pursuing a solution the equation is first simplified to the form

(20) $$|r'|^4 + a|r'|^2(r'\bar{a} + \bar{r}'a) + r'^2\bar{b} + \bar{r}'^2 b + |r'|^2 c + r'\bar{d} + \bar{r}'d + e = 0$$

where

(21) $$r = r' + a$$

(22) $$|r'|^2 = (|r|^2 - \bar{r}a - r\bar{a} + |a|^2)$$

$$r' = r - a, \quad \bar{r}' = \bar{r} - \bar{a}$$

$$r'^2 = r^2 - 2ar + a^2, \quad \bar{r}'^2 = \bar{r}^2 - 2\overline{ar} + \bar{a}^2)$$

(23) $$|r|^4 + 2|r|^2 r\bar{a} + 2|r|^2\bar{r}a = |r|^4 - 2|r|^2(\bar{r}a + r\bar{a}) + 2|r|^2|a|^2$$
$$+ \bar{r}^2 a^2 + 2|r|^2|a|^2 + r^2\bar{a}^2 - 2|a|^2(\bar{r}a + r\bar{a}) + |a|^4 + 2\bar{a}(r|r|^2 - |r|^2 a$$
$$- r^2\bar{a} + r|a|^2 - a|r|^2 + \bar{r}a^2 + r|a|^2 - a|a|^2) + 2a(\bar{r}|r|^2 - \bar{r}^2 a - |r|^2\bar{a}$$
$$+ \bar{r}|a|^2 - \bar{a}|r|^2 + \bar{r}|a|^2 + r\bar{a}^2 - \bar{a}|a|^2)$$

(24) $$|r|^4$$
$$r^2(\bar{b} - a^2) \qquad (\simeq \bar{\eta}r^2)$$
$$\bar{r}^2(b - a^2) \qquad (\simeq \eta\bar{r}^2)$$
$$|r|^2(-4|a|^2 + c) \quad (\simeq \xi|r|^2)$$
$$r\ (4\bar{a}|a|^2 - 2a\bar{b} + \bar{a}c + \bar{d}) \quad (\simeq \bar{f}r)$$
$$\bar{r}\ (4a|a|^2 - 2\bar{a}b + a\bar{c} + d) \quad (\simeq f\bar{r})$$
$$e = \Psi.$$

Finding the solution in this form is not pursued, rather a transformation to the variables (d,ϕ), is made.

$$d = |r|$$

$$\phi = \text{arc tan}\ \frac{r\ \text{imaginary}}{r\ \text{real}}$$

and $\quad r = d(\cos \phi + i \sin \phi)$

$$B = |\eta|$$

$$A = \text{arc tan}\ \frac{\eta\ \text{imaginary}}{\eta\ \text{real}}$$

$$F = |f|$$

$$C = \text{arc tan}\ \frac{f\ \text{imaginary}}{f\ \text{real}}$$

(25) $$\psi = d^4 + d^2 B(e^{2i\phi}e^{-iA} + e^{-2i\phi}e^{iA}) + d^2\xi + dF(e^{i\phi}e^{-iC} + e^{-i\phi}e^{iC}) + e$$

and since $\quad \cos \Theta = \frac{1}{2}(e^{i\Theta} + e^{-i\Theta})$

(26) $$\psi = d^4 + d^2(2B \cos(2\phi - A) + \xi) + 2dF \cos(\phi - C) + e.$$

Setting $$\phi' = \phi - \frac{1}{2}A$$

$$\omega = \frac{1}{2}A - C$$

(27) $$\psi = d^4 + d^2(2B \cos 2\phi' + \xi) + 2dF \cos(\phi' + \omega) + e = 0.$$

We look for points where ψ is stable with respect to both d and ϕ, for such are the minima we seek. By varying ψ with respect to both variables, we find points of zero variation.

(28) $$0 = 4d^3 + 2d(2B \cos 2\phi' + \xi) + 2F \cos(\phi' + \omega))$$

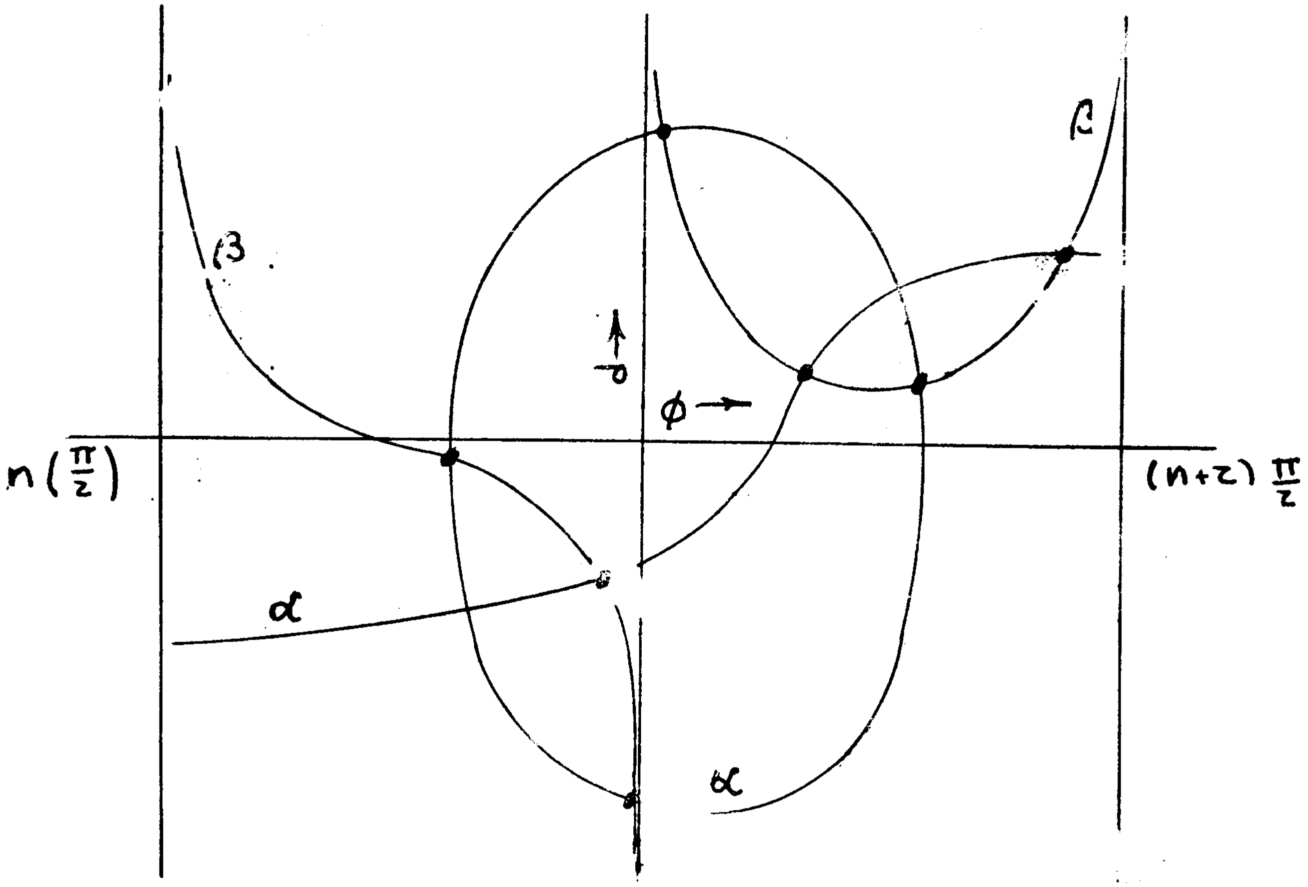
$$0 = -4d^2B \sin 2\phi' + 2dF \sin(\phi' + \omega)$$

(29) $$\alpha\!: \quad 0 = 2d^3 + d(2B \cos 2\phi' + \xi) + F \cos(\phi' + \omega)$$

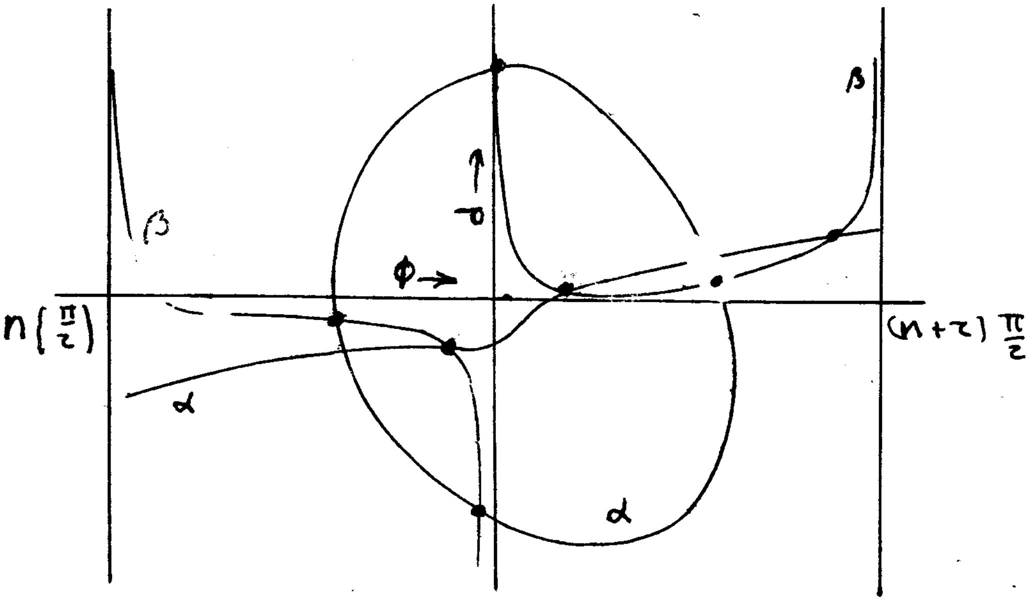
$$\beta\!: \quad 0 = 2dB \sin 2\phi' + F \sin(\phi' + \omega)$$

where these two curves intersect are the solutions to the system. We choose the extremum yielding the lowest value of ψ and with it correct Z. Thus we arrive at a new point, for which ψ is less than any previous value. The procedure should converge rapidly.

Solving α and β, then, is the primary complication. The graph of the function in the significant range $n(\frac{\pi}{2}) \le \not{\phi} \ge \le (n + 2)\frac{\pi}{2}$



with solution points indicated. By initially setting B large with respect to f and ξ , and by moving B to a canonical position, we obtain a graph whose solutions may be estimated closely by inspection.



Then, by changing B and ω by small increments, we can step back to their original values B_{real} and ω_{real} and follow the loci of all the solution points. Thus can all seven (possible) roots be found, and the best chosen.

Two things could go wrong. First, near the intersection of two loci, one locus might change course and start to follow the other. To handle this situation, we save past values of δ(d, ∅) and compute sin A, were A is the angle between the last and the present gradient at the point of solution. Should sin A increase sharply for two consecutive steps, the course is assumed changed and the tracing is restarted,this time with B and ω increments half as large.

A locus might also leave the real plane. Since the two points of intersection of two curves disappear together, these two loci are traced simultaneously, and should they approach each other so closely that their coincidence is imminent, they are discarded together. The algorithm uses Newton's generalized method to converge on solution points along the loci.

The equations

$$f(x_1, x_2, \ldots, x_n) = 0$$

$$g(x_1, x_2, \ldots, x_n) = 0$$

$$\ldots$$

$$h(x_1, x_2, \ldots, x_n) = 0$$

whose simultaneous solutions include $(x_{1s}, x_{2s}, \ldots, x_{ns})$ can be expanded about any point $0 = f(x_{1s}, x_{2s}, \ldots) \approx f(x_1, x_2, \ldots, x_n) + \frac{\partial f}{\partial x_1}(x_1 - x_{1s}) + \frac{\partial f}{\partial x_2}(x_2 - x_{2s}) + \ldots$

$$0 = g(x_{1s}, x_{2s}, \ldots) \approx g(x_1, x_2, \ldots) + \frac{\partial g}{\partial x_1}(x_1 - x_{1s}) + \ldots \; .$$

Under the circumstances that a set of simultaneous equations is to be solved and a good approximation to a solution is already known, the method of Newton, in a generalized form, can be used to improve the solution. The technique is to expand each of the functions about the approximate solution in a Taylor's expansion and then, discarding all but the linear terms of the expansions, proceed to solve the set of simultaneous equations that arise. Say, that it is desired to solve

(28) $$\alpha = 2d^3 + d(2B \cos 2\phi' + \xi) + F \cos(\phi' + \omega) = 0$$

(28a) $$\beta = 2dB \sin 2\phi' + F \sin(\phi' + \omega) = 0$$

we obtain

(29)
$$\text{d-d sol.} = \frac{\begin{vmatrix} -\alpha(d, \phi') & \frac{\partial\alpha}{\partial\phi'} \\ -\beta(d, \phi') & \frac{\partial\beta}{\partial\phi'} \end{vmatrix}}{\begin{vmatrix} \frac{\partial\alpha}{\partial d} & \frac{\partial\alpha}{\partial\phi'} \\ \frac{\partial\beta}{\partial d} & \frac{\partial\beta}{\partial\phi'} \end{vmatrix} = \text{Jacobian}}$$

(30)
$$\phi'\text{-}\phi' \text{ sol.} = \frac{\begin{vmatrix} \frac{\partial\alpha}{\partial d} & -\alpha(d, \phi') \\ \frac{\partial\beta}{\partial d} & -\beta(d, \phi') \end{vmatrix}}{\text{Jacobian}}$$

where

(31)
$$\frac{\partial\alpha}{\partial d} = 6d^2 + 2\beta \cos 2\phi' + \xi$$

$$\frac{\partial\alpha}{\partial\phi'} = -4dB \sin 2\phi' - F \sin (\phi' + \omega)$$

$$\frac{\partial\beta}{\partial d} = 2B \sin 2\phi'$$

$$\frac{\partial\beta}{\partial\phi'} = 4dB \sin 2\phi' + F \cos (\phi' + \omega)$$

Iteration of this procedure will quickly converge to an arbitrarily close approximation to the solution if the starting approximation was adequately accurate.

A similar procedure can be used to keep close to the locus of the solution of a set of equations,as the coefficients of the equations are changed. Actually the procedure that follows can be applied to the following of any contour of the solution space and is particularlized in this case to be the following of the locus of the solution.

For the finite difference stepping procedure,

(32)
$$\Delta d = \frac{\partial d}{\partial b}\Delta b + \frac{\partial d}{\partial \omega}\Delta\omega$$

$$\Delta\phi = \frac{\partial\phi}{\partial b}\Delta b + \frac{\partial\phi}{\partial\omega}\Delta\omega$$

where from $\alpha(d, \phi, B)$, $\alpha(d, \phi, \omega)$, $\beta(d, \phi, B)$, $\beta(d, \phi, \omega)$ we get by finding the coefficients of the variations

(33)
$$\frac{\partial\alpha}{\partial d}\cdot\frac{\partial d}{\partial B} + \frac{\partial\alpha}{\partial\phi}\frac{\partial\phi}{\partial B} + \frac{\partial\alpha}{\partial B} = 0$$

$$\frac{\partial\beta}{\partial d}\cdot\frac{\partial d}{\partial B} + \frac{\partial\beta}{\partial\phi}\frac{\partial\phi}{\partial B} + \frac{\partial\beta}{\partial B} = 0$$

$$\frac{\partial\alpha}{\partial d}\cdot\frac{\partial d}{\partial\omega} + \frac{\partial\alpha}{\partial\phi}\frac{\partial\phi}{\partial\omega} + \frac{\partial\alpha}{\partial\omega} = 0$$

$$\frac{\partial\beta}{\partial d}\cdot\frac{\partial d}{\partial\omega} + \frac{\partial\beta}{\partial\phi}\frac{\partial\phi}{\partial\omega} + \frac{\partial\beta}{\partial\omega} = 0\ .$$

Now setting

(34)
$$J = \begin{vmatrix} \frac{\partial\alpha}{\partial d} & \frac{\partial\alpha}{\partial\phi} \\ \frac{\partial\beta}{\partial d} & \frac{\partial\beta}{\partial\phi} \end{vmatrix} ,$$

gives

$$\Delta d = \left(\begin{vmatrix} -\frac{\partial\alpha}{\partial B} & \frac{\partial\alpha}{\partial\phi} \\ -\frac{\partial\beta}{\partial B} & \frac{\partial\beta}{\partial\phi} \end{vmatrix} \Delta B + \begin{vmatrix} -\frac{\partial\alpha}{\partial\omega} & \frac{\partial\alpha}{\partial\phi} \\ -\frac{\partial\beta}{\partial\omega} & \frac{\partial\beta}{\partial\phi} \end{vmatrix} \Delta\omega \right) / J$$

$$\Delta\phi = \left(\begin{vmatrix} \frac{\partial\alpha}{\partial d} & -\frac{\partial\alpha}{\partial B} \\ \frac{\partial\beta}{\partial d} & -\frac{\partial\beta}{\partial B} \end{vmatrix} \Delta B + \begin{vmatrix} \frac{\partial\alpha}{\partial d} & -\frac{\partial\alpha}{\partial\omega} \\ \frac{\partial\beta}{\partial d} & -\frac{\partial\beta}{\partial\omega} \end{vmatrix} \Delta\omega \right) / J$$

where $\frac{\partial\alpha}{\partial B} = 2d \cos 2\phi$

$$\frac{\partial\alpha}{\partial\omega} = -F \sin(\phi + \omega)$$

$$\frac{\partial\beta}{\partial B} = 2d \sin 2\phi$$

$$\frac{\partial B}{\partial\omega} = F \cos(\phi + \omega)$$

and $\Delta B = (B_{init} - B_{real})/N$

$$\Delta\omega = (\omega \text{ canon} - \omega \text{ real})/N$$

where N steps are taken to find the solution for the exact equation.

These stepping equations can then be used to find an approximation of the solution for the equations with modified coefficients. From the approximation it is then possible to find a solution to the desired accuracy by applying Newton's method. The equations can then be further modified and the process continued iteratively until the solutions for the exact equations are found.

APPENDIX II

The program operates as follows, then.

1) Initialize, read in values for S. The Z's are set initially to 1.

2) Compute $N = Z^+ - S$

$$k = \sum_i z_i \bar{z}_i$$

$$A = \sum_i z_i \bar{n}_i$$

$$\ell = \sum_i n_i \bar{n}_i$$

3) Compute $\text{grad } \psi = k\bar{N} + \ell Z - AN - \overline{AZ} = V$

Compute $C = \sum_i v_i \bar{z}_i$

$$D = \sum_i v_i \bar{z}_{i-1}$$

$$F = \sum_i v_i \bar{n}_i$$

$$E = \sum_i v_i \bar{n}_{i-1}$$

$$m = \sum_i v_i \bar{v}_i$$

$$B = \sum_i v_i \bar{v}_{i+1}$$

4) Compute $a_o = m - |B|^2$

5) Compute $\bar{a} = (m(C + E) - BD - \bar{B}F)/a_o$

$$\bar{b} = (CE - DF)/a_o$$

$$c = (m(k + \ell) + E\bar{C} + \bar{E}C - AB - \overline{AB} - |D|^2 - |F|^2)/a_o$$

$$\bar{d} = (kE - D\bar{A})/a_o$$

$$e = (k\ell - |A|^2)/a_o$$

6) Compute $b - a^2 = \eta$

$$c - 4|a|^2 = \xi$$

$$a\xi - 2\bar{a}b + d = f$$

These last 3, plus e, are the coefficients of ψ.

7) Compute $B = |\eta|$, $F = |f|$

Compute $A = \tan^{-1} \frac{I(\eta)}{R(\eta)}$

Compute $\omega = \tan^{-1} \frac{I(f)}{R(f)} - 1/2\, A$

8) Choose quadrants in which solutions will lie. There are two possible cases, depending on the amplitude of ω.

Do steps 9 - 18 for each of 4 pairs of loci.

9) Set initial values to B and ω.

Set initial approximation to (d,ϕ) for each locus of pair.

10) Set MPH = number of iterations.

Set $\Delta B = \frac{\text{Binitial - Breal}}{\text{MPH}}$

Set $\Delta\omega = \frac{\omega\text{canon. - }\omega\text{real}}{\text{MPH}}$

11) Converge on solution for initial B,ω.

Do steps 12 - 16 MPH times

Do steps 12 - 15 for each locus of pair.

12) Change B by ΔB, ω by $\Delta\omega$.

Step off $(\Delta d, \Delta\phi)$ for change in B,ω.

13) Compare $(\Delta d, \Delta\phi)$ to 3 previous values. If the sine of the angle of their change is large, go to step 14. Otherwise, to step 15.

14) If last value of sine was also too large, set MPH to 2 MPH+ 1, and go back to step 10. Otherwise, set $(\Delta d, \Delta\phi)$ to last value of $(\Delta d, \Delta\phi)$, substitute unconverged value of present point for converged value, and proceed.

15) Converge on new point (d, ϕ) by Newton's Method.

16) Test to determine if pair of solutions will disappear; if so, go back to step 9, for next pair of solutions.

17) Compute ψ for new solution. Determine if solutions found improve value of ψ. If so, solve new solutions and replace value of ψ with improved value.

18) Test if all pairs of solutions have been found. If not, change initial conditions and return to step 9.

19) Correct the root last recorded in step 18 to higher accuracy by Newton's Method.

20) Find real and imaginary parts of R' from solution of step 19.

21) Correct R' by translator a: $R = R' + \alpha$.

22) Correct Z. $Z = Z + RV$.

23) Test ψ. If $\psi > \epsilon$, go back to step 1.

24) $\psi \leq \epsilon$, set $\lambda = z_2 - s_1$

25) Reduce S by synthetic division

$$s_i = s_i + \lambda s_{i-1} \quad , \qquad (s_o = 1)$$

26) Test order of S to see if all eigenvalues have been found. If not, go to step 1.

APPENDIX III

```
C       FORTRAN PROGRAM TO FIND THE EIGENVALUES OF AN ARBITRARY MATRIX
        DIMENSION V(50,2),H(22),D(8,2),G(2,2,4),LO(2)
C       SUBROUTINE TO MULTIFLY TWO COMPLEX NUMBERS
        SUBROUTINE KMULT
        NONLOCAL V,MU,L1,L2,MUS,T
        DO 1 IC=1,2
        IS=3-IC
1       V(MU,IC)=V(MU,IC)+T'(V(L2,1)'V(L1,IC)+MUS'(IS-IC)'
      1 V(L2,2)'V(L1,IS))
        RETURN
        END
C       SUBROUTINE TO COMPUTE THE DIFFERENTIALS NEEDED IN
C       NEWTON AND OTHER PLACES
        SUBROUTINE DIFFER(N)
        NONLOCAL H,D,MUS,XSI
        H(22)=MUS'D(1,N)-H(17)
        H(8)=SINF(2'H(22))
        H(9)=COSF(2'H(22))
        H(10)=SINF(H(22)+H(6))
        H(22)=COSF(H(22)+H(6))
        H(1)=6'D(2,N)'D(2,N)+2'H(5)'H(9)+XSI
        H(2)=-4'D(2,N)'H(5)'H(8)-H(14)'H(10)
        H(3)=2'H(5)'H(8)
        H(4)=4'D(2,N)'H(5)'H(9)+H(22)'H(14)
        H(7)=H(1)'H(4)-H(2)'H(3)
        RETURN
        END
C       SUBROUTINE TO CONVERGE ON A SOLUTION POINT BY NEWTON'S METHOD
        SUBROUTINE NEWTON(N)
        NONLOCAL XSI,H,D,EPSI,ERSI
1       H(12)=D(2,N)'(2'D(2,N)'D(2,N)+
      1 H(5)'H(9)+XSI)+H(14)'H(22)
        H(13)=4'D(2,N)'H(5)'H(8)+H(14)'H(10)
        CALL DIFFER(N)
        DO 4 I=1,2
        H(I+10)=(3-2'I)'(H(13)'H(I)-H(12)'H(I+1))/H(7)
4       D(I,N)=D(I,N)-H(I+10)
        IF(H(12)/D(2,N)-ERSI) 2,2,1
2       IF(H(11)-EPSI) 3,3,1
3       RETURN
        END
        INPUT J,(V(I,1),V(I,2),I=1,J),EPSI,ERSI,UNIT,[illegible]
C       INITIALIZE
        H(20)=1000000000.
1       DO 2 I=1,J
        V(I+10,1)=1
2       V(I+10,2)=0
3       DO 4 I=21,49
        DO 4 IC=1,2
4       V(I,IC)=0
C       COMPUTE N'1
        DO 6 I=1,J
        DO 5 IC=1,2
5       V(I+20,IC)=V(I+11,IC)-V(I,IC)
C       COMPUTE k,A,1
        T=1
        MUS=1
```

```
          DO 6 K=0,2
          L2=K'(K-1)'5+10+I
          L1=K'(3-K)'5+10+I
          MU=K+41
 6        CALL KMULT
          DO 9 I=2,J
 C        COMPUTE GRADIENT V
          MU=I+30
          DO 8 K=1,2
          DO 8 MPH=0,1
          L1=I+K'10-MPH
          L2=44-K-MPH
          MUS=-MUS
          T=((MPH+K-1)'(MPH+K-4)-1)/V(43,1)
 8        CALL KMULT
 C        COMPUTE B,C,D,E,F,m
          T=1
          L1=I+30
          MU=43
          DO 9 K=10,30,10
          DO 9 MPH=0,1
          MU=MU+1
          L2=I+K-MPH
 9        CALL KMULT
 C        COMPUTE a0,c,e
          V(29,1)=V(48,1)'V(48,1)-V(49,1)'V(49,1)-V(49,2)'V(49,2)
          V(26,1)=V(48,1)'(V(41,1)+V(43,1))+2'(V(47,1)'V(44,1)+V(47,2)
        1 'V(44,2))-2'(V(42,1)'V(49,1)-V(42,2)'V(49,2))-V(45,1)'V(45,1)
        2 -V(45,2)'V(45,2)-V(46,1)'V(46,1)-V(46,2)'V(46,2)
          V(21,2)=V(41,1)'V(43,1)-V(42,1)'V(42,1)-V(42,2)'V(42,2)
 C        COMPUTE COMPLEX COEFFICIENTS a,b,d
          DO 17 IC=1,2
          IS=3-IC
          K=IS-IC
          V(27,IC)=V(48,1)'(V(44,IC)+V(45,IC))-V(49,IC)'(V(45,IC)-
        1 K'V(45,2)+V(46,1)'K+V(46,2))
          V(28,IC)=V(44,1)'V(47,IC)-V(45,IC)'V(46,IC)-K'(V(44,2)'V(47,IS)
        1 -V(45,2)'V(46,IS))
 17       V(25,IC)=V(41,1)'V(47,IC)+V(43,1)'V(44,IC)-V(42,IC)'(V(45,2)
        1 +V(46,1))+(V(46,2)+V(45,1)'K)
 C        DIVIDE ALL COEFFICIENTS BY a0
          DO 18 I=21,26
          DO 18 IC=1,2
 18       V(I,IC)=V(I,IC)/V(29,1)
 C        COMPUTE XSI
          XSI=V(28,1)-4'(V(27,1)'V(27,1)+V(27,2)'V(27,2))
 C        COMPUTE ETA AND f
          DO 19 IC=1,2
          IS=3-IC
          K=IS-IC
          V(23,IC)=K'(V(28,IC)+V(27,2)'V(27,IS))-V(27,1)'V(27,IC)
 19       V(22,IC)=K'(V(25,IC)-V(27,IC)'(XSI+2'(V(28,1)'K+V(28,2))))
          OUTPUT (V(I,1),V(I,2),I=21,49)
 C        BRANCH OF PROGRAM TO COMPUTE BEST VALUE FOR R
 C        INITIALIZE
          MUS=-1
          M(11)=1
```

```
C     COMPUTE A,B,OMEGA,AND F
      DO 30 I=22,23
      H(I-7)=.5'(ATANF(V(I,2)/V(I,1))+SIGNF(H(11),V(I,1))-1)'
     1 SIGNF(H(11),V(I,2)))'1.57079632
   30 H(I-9)=SQRTF(V(I,1)'V(I,1)+V(I,2)'V(I,2))
      H(21)=2'H(13)-H(15)
      H(17)=-6.38318528
C     FIND THE QUADRANT IN WHICH (-OMEGA) LIES
   32 H(17)=H(17)+1.57079632
      MUS=-MUS
      IF(H(21)+H(17)) 32,34,33
C     IF OMEGA IS A MULTIPLE OF PI/2,IT WAS CHANGED SLIGHTLY
C     TO SIMPLIFY COMPUTATION
   34 H(21)=H(21)+.157079632
   33 IF(MUS) 36,35,35
   35 MUS=-MUS
      H(17)=H(17)-1.57079632
C     COMPUTE APPROXIMATIONS OF SOLUTIONS FOR FIRST PAIR
   36 D(5,1)=0.157079632
      D(5,2)=1.41371669
      D(6,1)=H(14)
      D(6,2)=-H(14)/H(16)
C     MAIN LOOP:ONE PASS FOR EACH PAIR OF SOLUTIONS
      DO 37 IS=1,4
C     AT THE BEGINNING OF EACH PASS,MPH IS SET EQUAL TO MANY.
C     IF IT BECOMES NECESSARY DURING A PASS, THE VALUE WILL BE INCREASED
      MPH=MANY
      MU=-1
C     SET B TO ITS INITIAL VALUE
   57 H(11)=(H(14)+ABSF(XSI))'INIT
C     AFTER THE SECOND PASS, THE INITIAL VALUES ARE SLIGHTLY CHANGED
      IF (IS-3)54,53,54
   53 D(6,1)=ABSF(XSI-H(11))
      D(5,2)=0.78539806
C     COMPUTE DELTA OMEGA AND DELTA B
   54 H(19)=ABSF(1.41371669-ABSF(H(17)+H(21)))/MPH
      H(18)=(H(11)-H(16))/MPH
      H(5)=H(11)
      H(6)=-1.41371669
C     SET dAND PHI TO THEIR INITIAL APPROXIMATIONS AND
C     CONVERGE ON SOLUTIONS, USING NEWTON'S METHOD
      DO 38 IC=1,2
      DO 56 I=1,2
      D(I,IC)=MU'D(I+4,IC)
C     FOR EACH PASS, THE INITIAL APPROXIMATIONS ARE THE NEGATIVES
C     OF THOSE OF THE LAST PASS
   56 D(I+4,IC)=D(I,IC)
   38 CALL NEWTON (IC)
C     LOOP TO STEP TO DESIRED SOLUTION FOR EACH POINT
      DO 40 L1=1,MPH
      DO 41 IC=1,2
      CALL DIFFER (IC)
C     COMPUTE DELTA d AND DELTA PHI
      DO 58 I=1,2
      G(I,IC,1)=(3-2'I)'(2'D(2,IC)'(H(9)'H(I+2)-H(8)'H(I))'H(18)
     1 -H(21)'(H(22)'H(I)-H(10)'H(I+2))'H(19))/H(7)
C     THE COMPARISON FOR SINES IS LOCKED OUT OF THE FIRST 3 PASSES
      IF(L1-4) 63,62,62
```

```
COMPUTE THE SINE OF THE ANGLE OF CHANGE FOR THE LAST
C         THREE DELTAS
62        H(11)=ABSF(G(I,IC,4)'G(I,IC,2)-G(I,IC,3)'G(I,IC,1))
         1 /SQRTF((G(I,IC,1)'G(I,IC,1)+G(I,IC,3)'G(I,IC,3))'
         2 (G(I,IC,2)'G(I,IC,2)+G(I,IC,4)'G(I,IC,4)))
C         SINE TEST-IS THE LOCUS ORDERLY ENOUGH?
          IF (ABSF(H(11)/D(I+6,IC))-3) 63,63,61
C         IF SO,SET LO TO 1
63        LO(IC)=1
C         REPLACE THE OLD COMPUTED SINE BY THE NEW ONE
          D(I+6,IC)=H(11)
C         INCREMENT d AND PHI AND STORE THE UNCONVERGED VALUE
65        D(I,IC)=D(I,IC)+G(I,IC,1)
          D(I+2,IC)=D(I,IC)
C         MOVE THE PAST DELTAS BACK ONE TIME UNIT
          DO 58 K=1,3
          L2=5-K
58        G(I,IC,L2)=G(I,IC,L2-1)
C         CONVERGE TO POINT ON LOCUS
          CALL NEWTON (IC)
C         ON THE FIRST PASS,THERE IS ONLY ONE SOLUTION BEING FOLLOWED
          IF (IS-1) 41,90,41
C         THE TEST FOR SINE FAILED.SET TRIP FOR NEXT PASS,SET PRESENT DELTA TO LAST
C         DELTA,AND INCREENT THE UNCONVERGED RATHER THAN THE  CONVERGED
C         VALUE OF THE PRESENT POINT.
61        IF(LO(IC)) 92,64,64
64        G(I,IC,1)=G(I,IC,2)
          D(I,IC)=D(I+2,IC)
          LO(IC)=-1
          GOTO 65
C         IF THE SINE WAS TOO LARGE FOR TWO SUCCESSIVE STEPS,WE
C         SET MPH TO A MUCH LARGER VALUE AND START AGAIN
92        MPH=2'MPH+1
          MU=1
          GOTO 57
41        CONTINUE
C         TEST IF THE TWO POINTS ARE TOO CLOSE,AND MUST BE DISCARDED
          DO 48 IC=1,2
          DO 48 I=1,2
          IF(ABSF((D(I,1)-D(I,2))/G(I,IC,1))-3) 90,90,48
48        CONTINUE
          GO TO 37
C         INCREMENT B AND OMEGA
90        H(5)=H(5)-H(18)
40        H(6)=H(6)-H(19)
C         TWO SOLUTIONS HAVE BEEN FOUND.COMPUTE PSI
91        DO 37 IC=1,2
          CALL DIFFER(IC)
          H(8)=D(2,IC)'D(2,IC)'(D(2,IC)'D(2,IC)+H(16)'
        1 H(9)+XSI)+H(14)'H(22)+V(21,2)
C         IS THE NEW PSI BETTER THAN THE OLD?
          IF(H(20)-H(8)) 37,37,51
C         IF SO, REPLACE THE OLD VALUES WITH THE NEW
51        D(1,IC)=D(I,IC)+H(15)
          V(23,1)=D(2,IC)'COSF(D(1,IC))
          V(23,2)=D(2,IC)'SINF(D(1,IC))
          H(20)=H(8)
```

```
C       IF PSI IS SMALL ENOUGH,JUMP OUT OF ITERATION
        IF (H(20)-EPSI) 55,55,37
37      CONTINUE
55      DO 20 IC=1,2
20      V(22,IC)=V(23,IC)+V(27,IC)
        DO 13 IC=1,2
        IS=3-IC
        DO 13 I=1,J
C       CHANGE Z BY Z=Z+R'V
13      V(I+10,IC)=V(22,1)'V(I+30,IC)+(IS-IC)'V(22,2)'V(I+30,IS)
        IF(H(20)-EPSI) 14,14,3
14      OUTPUT V(22,1),V(22,2),H(20)
C       AN EIGENVALUE HAS BEEN FOUND.REDUCE S BY SYNTHETIC DIVISION
        V(1,1)=1
        L2=22
        MUS=-1
        DO 15 MU=2,J
        L1=I-1
15      CALL KMULT
        J=J-1
        V(1,1)=0
C       HAVE ALL EIGENVALUES BEEN FOUND? IF NOT,GO BACK TO BEGINNING
        IF(J) 1,16,1
16      STOP
        END
        END
```